A Simple Guide to Using Multilevel Models for the Evaluation of Program Impacts

Gustavo Angeles and Thomas A. Mroz

March 2001

MEASURE Evaluation

Carolina Population Center
University of North Carolina
at Chapel Hill
123 W. Franklin Street
Suite 304
Chapel Hill, NC 27516
Phone: 919-966-7482
Fax: 919-966-2391
measure@unc.edu
www.cpc.unc.edu/measure

<u>Collaborating Partners:</u>

Macro International Inc.
11785 Beltsville Drive
Suite 300
Calverton, MD 20705-3119
Phone: 301-572-0200
Fax: 301-572-0999
measure@macroint.com

John Snow Research and Training Institute
1616 N. Ft. Myer Drive
11th Floor
Arlington, VA 22209
Phone: 703-528-7474
Fax: 703-528-7480
measure_project@jsi.com

Tulane University
1440 Canal Street
Suite 2200
New Orleans, LA 70112
Phone: 504-584-3655
Fax: 504-584-3653
measure2@tulane.edu

<u>Funding Agency:</u>

Center for Population, Health and Nutrition
U.S. Agency for
International Development
Washington, DC 20523-3600
Phone: 202-712-4959

WP-01-33

The research upon which this paper is based was sponsored by the MEASURE *Evaluation* Project with support from the United States Agency for International Development (USAID) under Contract No. HRN-A-00-97-00018-00.

The working paper series is made possible by support from USAID under the terms of Cooperative Agreement HRN-A-00-97-00018-00. The opinions expressed are those of the authors, and do not necessarily reflect the views of USAID.

The working papers in this series are produced by the MEASURE *Evaluation* Project in order to speed the dissemination of information from research studies. Most working papers currently are under review or are awaiting journal publication at a later date. Reprints of published papers are substituted for preliminary versions as they become available. The working papers are distributed as received from the authors. Adjustments are made to a standard format with no further editing.

A listing and copies of working papers published to date may be obtained from the MEASURE *Evaluation* Project at the address listed on the back cover.

<u>Other MEASURE *Evaluation* Working Papers</u>

WP-01-32: The Effect of Structural Characteristics on Family Planning Program Performance in Côte d'Ivoire and Nigeria (Dominic Mancini, Guy Stecklov and John F. Stewart)

WP-01-31: Socio-Demographic Context of the AIDS Epidemic in a Rural Area in Tanzania with a Focus on People's Mobility and Marriage (J. Ties Boerma, Mark Urassa, Soori Nnko, Japheth Ng'weshemi, Raphael Isingo, Basia Zaba, and Gabriel Mwaluko)

WP-01-30: A Meta-Analysis of the Impact of Family Planning Programs on Fertility Preferences, Contraceptive Method Choice and Fertility (Gustavo Angeles, Jason Dietrich, David Guilkey, Dominic Mancini, Thomas Mroz, Amy Tsui and Feng Yu Zhang)

WP-01-29: Evaluation of Midwifery Care: A Case Study in Rural Guatemala (Noreen Goldman and Dana A. Glei)

WP-01-28: Effort Scores for Family Planning Programs: An Alternative Approach (John A. Ross and Katharine Cooper-Arnold)

WP-00-27: Monitoring Quality of Care in Family Planning Programs: A Comparison of Observation and Client Exit Interviews (Ruth E. Bessinger and Jane T. Bertrand)

WP-00-26: Rating Maternal and Neonatal Health Programs in Developing Countries (Rodolfo A. Bulatao and John A. Ross)

WP-00-25: Abortion and Contraceptive Use in Turkey (Pinar Senlet, Jill Mathis, Siân L. Curtis, and Han Ruggers)

WP-00-24: Contraceptive Dynamics among the Mayan Population of Guatemala: 1978-1998 (Jane T. Bertrand**,** Eric Seiber and Gabriela Escudero)

WP-00-23: Skewed Method Mix: a Measure of Quality in Family Planning Programs (Jane T. Bertrand, Janet Rice, Tara M. Sullivan & James Shelton)

WP-00-22: The Stymied Contraceptive Revolution in Guatemala (Roberto Santiso G. and Jane T. Bertrand)

WP-00-21: The Impact of Health Facilities on Child Health (Eric R. Jensen and John F. Stewart)

WP-00-20: Effort Indices for National Family Planning Programs, 1999 Cycle (John Ross and John Stover)

WP-00-19: Evaluating Malaria Interventions in Africa: A Review and Assessment of Recent Research (Thom Eisele, Kate Macintyre, Erin Eckert, John Beier, and Gerard Killeen)

WP-00-18: Monitoring the AIDS epidemic using HIV prevalence data among young women attending antenatal clinics: prospects and problems (Basia Zaba, Ties Boerma and Richard White)

WP-99-17: Framework for the Evaluation of National AIDS Programmes (Ties Boerma, Elizabeth Pisani, Bernhard Schwartländer, Thierry Mertens)

WP-99-16: National trends in AIDS knowledge and sexual behaviour in Zambia 1996-98 (Charles Banda, Shelah S. Bloom, Gloria Songolo, Samantha Mulendema, Amy E. Cunningham, J. Ties Boerma)

WP-99-15: The Determinants of Contraceptive Discontinuation in Northern India: A Multilevel Analysis of Calendar Data (Fengyu Zhang, Amy O. Tsui, C. M. Suchindran)

WP-99-14: Does Contraceptive Discontinuation Matter?: Quality of Care and Fertility Consequences (Ann Blanc, Siân Curtis, Trevor Croft)

WP-99-13: Socioeconomic Status and Class in Studies of Fertility and Health in Developing Countries (Kenneth A. Bollen, Jennifer L. Glanville, Guy Stecklov)

WP-99-12: Monitoring and Evaluation Indicators Reported by Cooperating Agencies in the Family Planning Services and Communication, Management and Training Divisions of the USAID Office of Population (Catherine Elkins)

WP-98-11: Household Health Expenditures in Morocco: Implications for Health Care Reform (David R. Hotchkiss, Zine Eddine el Idriss, Jilali Hazim, and Amparo Gordillo)

WP-98-10: Report of a Technical Meeting on the Use of Lot Quality Assurance Sampling (LQAS) in Polio Eradication Programs

WP-98-09: How Well Do Perceptions of Family Planning Service Quality Correspond to Objective Measures? Evidence from Tanzania (Ilene S. Speizer)

WP-98-08: Family Planning Program Effects on Contraceptive Use in Morocco, 1992-1995 (David R. Hotchkiss)

WP-98-07: Do Family Planning Service Providers in Tanzania Unnecessarily Restrict Access to Contraceptive Methods? (Ilene S. Speizer)

WP-98-06: Contraceptive Intentions and Subsequent Use: Family Planning Program Effects in Morocco (Robert J. Magnani)

WP-98-05: Estimating the Health Impact of Industry Infant Food Marketing Practices in the Philippines (John F. Stewart)

WP-98-03: Testing Indicators for Use in Monitoring Interventions to Improve Women's Nutritional Status (Linda Adair)

WP-98-02: Obstacles to Quality of Care in Family Planning and Reproductive Health Services in Tanzania (Lisa Richey)

WP-98-01: Family Planning, Maternal/Child Health, and Sexually-Transmitted Diseases in Tanzania: Multivarate Results using Data from the 1996 Demographic and Health Survey and Service Availability Survey (Jason Dietrich)

A Simple Guide to Using Multilevel Models for the Evaluation of Program Impacts

Gustavo Angeles
Carolina Population Center

Thomas A. Mroz
Department of Economics and the Carolina Population Center

MEASURE Evaluation
University of North Carolina, Chapel Hill

March 13, 2001

This is a preliminary version. Andrew Dyke and Arthur Sinko provided valuable research assistance. We thank Guang Guo and David Guilkey for valuable comments on this paper.

Introduction

The purpose of this essay is to help researchers investigating the impacts of health, family planning, and nutrition programs understand the importance and relevance of using multilevel analysis in their empirical evaluations of the programs' impacts. The discussion first defines what it means to have a multilevel model, and it then turns to an examination of the statistical properties of estimators when one has a hierarchical structure. Throughout the essay we illustrate the basic points through the use of Monte Carlo experiments, where we simulate data and outcomes according to known and exact rules. After simulating data, we use a variety of estimation approaches to estimate the underlying relationships in the simulated data. Since we know the "true" way the "world" operates in these experimental settings, these Monte Carlo experiments allow us to evaluate how well particular statistical procedures can uncover the "true" form of the statistical relationship. Based on these Monte Carlo experiments and some direct comparisons of the statistical properties of the various estimators that we consider, we present a set of recommended approaches for using multilevel data to assess the overall effectiveness of programs.

We focus our analysis on simple multilevel models where the effects of observed covariates are fixed and do not vary across units of the hierarchical structure. The residual term in a linear regression model possibly has a simple hierarchical structure. Our primary concern is how well various estimators measure the impacts of observed covariates on outcomes of interest. We focus on unbiasedness of the point estimators, precision of the estimators, and the ability of the point and standard error estimators to provide unbiased hypothesis tests. For our evaluations we focus on only simple linear regression models with continuous outcomes estimated by ordinary least squares (OLS) and on simple, two-level maximum likelihood estimation models.

Given this scope, the essay reaches three main conclusions. First, if the data do have a multilevel error structure and one fails to account for this in the estimation of standard errors of estimates, one can dramatically overstate the significance of the estimated statistical relationships. In particular, a researcher who fails to use procedures that adjust estimated standard errors for the multilevel error structure would "uncover" statistically significant relationships when they do not exist. To obtain correct statistical inferences, one need not use complete multilevel modeling approaches. Instead, statistical procedures that ex post account for the clustering in the data when calculating standard errors will provide correct standard errors. Second, there typically is little efficiency loss in the estimation of the impact of a community-level variable on individual-level outcomes if one ignores the multilevel error structure and uses Ordinary Least Squares procedures to estimate the impacts of covariates on the outcome of interest. There can, however, be sizable increases in efficiency for estimators of the impacts of the individual-level variables, but these effects are typically of less interest in program evaluation studies. The third conclusion is more tentative than the first two. It deals with problems that one can encounter with multilevel models when one incorrectly assumes a simple linear relationship when the true relationship is nonlinear. In particular, if one imposes incorrectly a simple linear specification for the observed regressors when there really is a more complex function describing mean effects, then it is possible to incorrectly "uncover" a multilevel error structure when one does not exist. Taken as a whole these conclusions suggest that a fruitful estimation approach in practice would be to rely on simple estimation procedures like ordinary least squares, adjust the estimated standard errors to account for the possible multilevel error structures, and examine whether nonlinear relationships might better describe the data than simple linear effects. After a thorough examination of the empirical relationship with simple models and adjusted standard errors, one could then use more detailed multilevel models to obtain more precise estimators.

Heuristic Description of Multilevel Models

Multilevel models are used when the outcome of interest, and its observed and unobserved determinants, have an hierarchical structure. By an hierarchical structure, we mean that there are important factors influencing decisions and outcomes that arise from a variety of levels of aggregation or observation. For example, whether individuals use contraception might depend on whether there are easily accessible clinics in the community where they live where they can receive family planning counseling and contraceptives. The presence of such clinics, of course, could influence the contraceptive choice of many individuals living in the same community. Each clinic is available to more than one individual, and this gives rise to the multilevel structure of observed determinants of contraceptive use.

Typically the outcome of interest takes place at an individual level, and this usually is referred to as the lower- or micro-level outcome. In analyses with more than two levels, this is called the level-one outcome. These lower level, individual outcomes are usually influenced in part by individual, micro-level characteristics. In the family planning literature, for example, a woman's age and education and measures of her wealth have all been shown to have important effects on an individual's use of contraceptives (Gertler and Molyneaux, 1994; Guilkey and Cochrane, 1995; Guilkey and Jayne, 1997). Measures of whether there is a family planning clinic providing information about contraceptives in a woman's village is a higher-level, or macro-level, determinant of an individual's contraceptive use. Presumably, the characteristics of the clinic have somewhat similar effects on all individuals residing within the same community. These varying levels of outcomes and determinants, i.e., at the individual, family, community or regional level, give rise to the hierarchies analyzed with multilevel models. Kreft and de Leeuw (1998) provide an excellent introductions to multilevel models, and Goldstein (1995) and Byrk and Raudenbush (1992) present more advanced treatments of these modeling approaches.

What distinguishes the hierarchy in these types of analyses is the fact that some characteristics from a higher level also influence the lower-level outcomes. Researchers have found that food prices, for example, can influence whether a couple practices contraception (Stewart et al., 1991; Rous, 2001). High prices might indicate food shortages, or that it would be expensive to raise children, and couples might tend to be more likely to attempt to limit fertility when food prices are high than when food prices are low. Food prices, like many other contraceptive determinants, vary across communities but individuals within a single community all face the same level of food prices. Food prices and other variables specific to a higher, more aggregate level are higher-level determinants of individual-level contraceptive use. There can also be unobserved or unmeasured factors at the higher level that influence the lower-level outcomes. Such unmeasured factors give rise to multilevel error structures; these are discussed extensively in the following sections.

Family planning clinics are located within communities, and the availability of these sources for contraceptives and of contraceptive knowledge can affect whether individuals adopt family planning (Tsui, 1985; Tsui and Ochoa, 1992; Bollen, Guilkey and Mroz, 1995; Thomas and

Maluccio, 1995; Guilkey and Jayne, 1997; Angeles et al., 2001). Often such higher level determinants of contraceptive use are observed, and researchers usually include these measures as explanatory variables in their empirical analyses when they are available. At a basic level, there is nothing special about these higher level determinants that distinguishes them from individual characteristics like age and education. One can readily incorporate observed community-level characteristics along with observed individual-level characteristics as determinants of individual-level behaviors. The fact that these higher level characteristics do not differ within groups of individuals is, for the most part, irrelevant in the interpretation of impacts of observed covariates on individual-level outcomes.

What is important to recognize about the impacts of higher level factors on lower level behaviors is the fact that all individuals who face an identical higher level factors experience similar impacts from these higher level factors. All individuals in the same community, for example, would have the same clinic available to them; anything idiosyncratic about that particular clinic will have roughly the same impact on everyone within the community. Similarly, if there are unobserved or unmeasured community factors that influence the behaviors of individuals within each community, then there will be correlations of individual-level outcomes within each community after controlling for the observed individual-level and community-level covariates.

Statistical Consequences of Multilevel Models

It is the presence of unobserved or unmeasured higher level characteristics that makes it important for a researcher to adjust her analysis to accommodate and recognize the multilevel structure. Since one cannot control for these unmeasured community characteristics, their impacts on an outcome of interest are represented through their becoming a part of the "error term" in a statistical model. Consider, for example, an individual-level regression analysis using data on individuals within each of many communities. In this instance any unobserved or unmeasured community-level factors that have an impact on the outcome of interest would have an impact on the outcomes for all individuals within the community. This means that the error terms for individuals within each community could be correlated. Any unobserved community characteristic that influences one community member to have a high value for the outcome of interest would likely result in other members of the same community having similarly high values of the outcome of interest.

Such error correlations among individuals within communities imply that the standard statistical assumption that different observations have independent residuals will be violated. This correlation of the individual-level residuals within a community gives rise to several important statistical considerations. First, some of the desirable statistical properties of estimators rely upon assumptions of independent residuals; such an assumption is clearly incorrect when there are unobserved community-level factors influencing behaviors of individuals within communities. For example, the Gauss-Markov Theorem states that the ordinary least squares (OLS) estimator is the best linear estimator within the class of unbiased estimators. This optimality implication depends crucially upon the residuals being uncorrelated across all observations. In the presence of unobserved community-level effects, one can usually define a better unbiased, linear estimator than ordinary least squares, where in this context a better estimator means one that provides more accurate parameter estimates. Either a generalized least squares (GLS) estimator or a maximum likelihood, multilevel error components estimator would typically provide more accurate estimates than OLS in these types of situations.

It is important to recognize, however, that some key properties of commonly used estimators do not depend on error terms being independent across observations. Most importantly, whether or not the OLS estimator is unbiased or consistent does not depend upon residuals being uncorrelated. Nor do the unbiasedness and consistency of the OLS estimator require that residuals for all observations have the same variance. All that the OLS estimator requires for it to be an unbiased estimator (i.e. correct on average) is that the explanatory variables at both the micro and the macro levels be uncorrelated with the residuals. The error terms as well as the explanatory variables then can have both micro-level and macro-level components, and the estimator will still be unbiased. This is not a restrictive assumption for the class of estimation problems we consider. In fact, every one-equation estimator for a multilevel model requires at least this basic assumption that the explanatory variables are uncorrelated with the residuals in order for the estimator to be unbiased or consistent. The presence of multilevel or hierarchical unobserved factors does not lead to biased estimators of the effects of individual- or community-level factors on the individual-level outcomes.[1]

While the presence of correlated residuals does not result in a bias of the point estimates from these OLS estimators, estimators of standard errors, confidence intervals, and statistical significance will be biased and incorrect unless one explicitly recognizes the correlated residuals

---

[1]A companion paper to this study explores the supposed bias in discrete outcome models in the presence of multilevel error structures that have been reported in several published papers (Mroz, 2001). In most instances, the supposed biases are a result of the authors of these papers failing to recognize that coefficients in discrete outcome models have substantively different interpretations than regression models with continuous outcomes.

when constructing these measures. Consider, for example, a sample of 1,000 communities where the data set contains two individuals within each community. This yields a total individual-level sample size of 2,000 observations. We assume that observations are independent across the 1,000 communities. The typically used standard error estimators, such as those reported by OLS regression procedures, assume that there are 2,000 individual-level observations with uncorrelated error terms.

Now consider an extreme case where the two residuals for the two individual-level observations within each community are identical and where there are only community-level explanatory variables. In this specification no new information is provided by the second observation within the community as the second observation is identical to the first. One would obtain exactly the same parameter estimates if one uses only a randomly selected "first" observation in the community or if one uses only the "second" observation in the community or if one pools all 2,000 individual-level observations together. All that matters in this hypothetical example is that one has at least one observation from each community[2].

Since using any 1,000 independent observations (i.e., any one observation from each of the 1,000 communities) would yield exactly the same (i.e., identical) estimates as using all 2,000 observations, it cannot be the case that using the larger data set provides more accurate information. In this example, because of the perfect correlation of error terms and regressors within communities (and, hence, of outcomes), all of the relevant information is carried by 1,000 observations. Having an additional individual per community does not provide any new

[2]To simplify this discussion, we implicitly assume that if one uses more than one observation per community, then one uses exactly the same number of observations from all communities. I.e., the data set is always rectangular.

information. Parameter estimates cannot become more precise if one uses more than one individual per community instead of only one person per community, as the extra observations in this hypothetical example provide absolutely no new information. Yet the usual OLS formulae for calculating standard errors of the estimators were derived under the assumption that each additional observation provides new, independent information about the relationship between the outcome and the regressors.

In this example it is straightforward to work out the consequences of the error terms being perfectly correlated for observations within the same community. To simplify the discussion, suppose that there is only one explanatory variable. Recall that in this example the explanatory variable is constant for all observations within the community. Using only the first observation in each of the 1,000 communities, the true standard error of the least squares estimator under standard assumptions would be:

$$se(b_1) = \sqrt{Var(b_1)} = \sqrt{\frac{s^2}{\sum_{j=1}^{1000} (x_j - \bar{x})^2}}$$

where $\bar{x}$ is the mean of the explanatory variable across the 1,000 communities. This estimation procedure, based on the 1,000 observations, incorporates all of the information contained in the 2,000 observations; the additional 1,000 observations merely replicate the first 1,000 observations. If one uses all 2,000 observations the OLS point estimate would not change at all.

The naive, simple OLS estimator of the standard error of this estimator with 2 observations per community would be:

$$se_{naive,OLS}(b_2) = \sqrt{Var_{naive,OLS}(b_2)} = \sqrt{\frac{s^2}{\sum_{j=1}^{1000}\sum_{i=1}^{2}(x_j - \bar{x})^2}} = \sqrt{\frac{s^2}{2\sum_{j=1}^{1000}(x_j - \bar{x})^2}} = \frac{1}{\sqrt{2}}\,se(b_1)\ .$$

The true standard error of this estimator, however, must be exactly the same as the standard error for the estimator that uses only one observation per community, i.e., $se_{true}(b_2) \equiv se(b_1)$. The simple OLS formula for the standard error, based on the presumption that all observations are independent, provides standard errors that are smaller than the true standard errors by a factor of

$\frac{1}{\sqrt{2}} \approx \frac{1}{1.41}$. T-statistics calculated using the incorrect, simple OLS standard error will be 1.41

times larger than the true value of t-statistic (as calculated with the true standard error). By using this incorrect standard error estimator, in this instance, all calculated t-statistics will be measured as 41% higher than they should be. One will too frequently reject true null hypothesis, and all confidence intervals will be too short by a factor or 41%.

The use of the typical OLS standard error formulae is clearly incorrect in this example. There is absolutely no new information being used in the estimation with the 2,000 individuals that was not contained in the sample of 1,000 observations with one observation per community. This example is, of course, extreme. But it does make an important point. If there is a significant level of correlation of residuals within each community, then additional observations for each community do not provide as much “new” information to the estimation procedure as the first observation in the community provided. Naive, simple standard error estimators impose the incorrect assumption that each additional observation within a community provides just as much new information as the first observation in the community provided. The above example clearly demonstrates the potential for simple estimation procedures to yield incorrect standard errors and consequently incorrect inferences when one relies on standard error estimators that require independent observations.

When the error correlations are less than one, or if there are explanatory variables that vary among the lower level units within the higher level units, then the additional observations on level-one units can provide some new information that is not included in the first level-one observation for each higher level unit. In the general case, there will be somewhat less new information from each additional level-one observation within each higher level unit, so the naive standard error estimators will yield incorrect standard errors of the estimates. But these additional observations do provide some new information, so one might be able to obtain more efficient estimators by using all of the observations that are available when estimating regression functions. However, regardless of whether one uses point estimators that make efficient use of all of the information in a sample, it will be necessary to use procedures that recognize and control for the dependence of error terms within each higher level unit when calculating standard errors of the estimates and test statistics.

A Simple Statistical Framework

While the preceding heuristic discussion lays out some of the most important shortcomings of naive analyses with multilevel data, it is important to have a precise statistical formulation for addressing many of the issues one encounters in analyses of multilevel data. In this section we introduce some of the notation that we will use throughout our evaluation of approaches to use when one has access to multilevel data. Throughout this discussion we will use uppercase Roman letters to represent observable random variables and lowercase Roman letters to represent actual realizations of these random variables. Random variables labeled Y will stand for outcomes, and X will stand for observed explanatory variables. For most of this analysis it is not important whether the explanatory variables are fixed or random.

The main focus of this examination of multilevel models is on two-level models, where we call arbitrarily the higher level (level two) a community and the lower level (level one) an individual within the community. We will use subscripts c to indicate community and the subscript i to refer to an individual within a community. Following this notation, Y would be a random variable indicating an outcome, and y(i,c) would be the observed outcome for the i^{th} individual in the c^{th} community. We will assume that there are J communities (i.e., c=1,2,...,J) and that there are N_c individuals in community c (i.e., i=1,2,...,N_c within each community c). There are a total of $N^* = \sum_{c=1}^{J} N_c$ observations. For part of the discussion and analysis we will assume that there is an identical number of individual observations within each community. In this case, we will assume that N_c = N for all communities, and there will be a total of N·J individual-level outcomes that are observed in the data $(N^* = \sum_{c=1}^{J} N_c = \sum_{c=1}^{J} N = N \cdot J)$.

We differentiate between two broad categories of observed explanatory variables: community-level explanatory variables and individual-level explanatory variables. Community-level explanatory variables take on the same value for all individuals within each community c, while individual-level explanatory variables usually will differ among individuals within a community. An example of a community-level variable might be the presence of a health clinic within a community or the level of per capita expenditures on family planning programs within a community. These measures would not vary across individuals within each community while they would vary across communities. We use the symbol $X_c(c)$ to denote the value of the community-level variable in community c.

We consider two different types of individual-level explanatory variables. The first type contains variables that are correlated across individuals within communities. Often these are correlated with the observed community-level variable. We use the symbol $X_{IC}(i,c)$ to denote those individual-level variables that are correlated at the community level. An example of an individual-level explanatory variable that could be correlated among individuals within a community might be an individual's level of education. To see a source of the within community correlation, consider a study of the impact of community health clinics and mother's education on a child's health in a developing country. The quality of schools would most likely differ across communities, and one might expect that those individuals who live in areas with better schools would have stayed in school longer than those individuals who lived in areas with poor schools. If it were the case that communities with health clinics also tended to have better (worse) schools, then those with higher levels of schooling would tend to be concentrated in communities with (without) health clinics. This gives rise to a correlation between the level-two covariate, the presence of a health clinic, and the level-one covariate, the individual's education attainment. Even if the health clinics and good schools were not related, there would still be community-level correlation of mothers' educational attainments as long as there were differences across communities in individuals' access to good education programs and individuals made schooling decisions that take school quality into account.

The second type of individual-level variable encompasses those variables that are independent across individuals within a each community and uncorrelated with any of the community-level measures. We denote these by $X_I(i,c)$. Independent individual-level variables do not have any community-specific component. As an example of such independent variables, consider a case where the age distribution within each community is the same for all communities. In this instance the age of randomly selected individuals within each community would be such an independent individual-level variable. But if some communities tend systematically to have older individuals while others tend to have younger individuals, then the age of independently selected individuals would not be an independent individual-level variable; in this instance it would be an individual-level variable that is correlated among individuals within a community.

The unobserved or unmeasured determinants of an outcome of interest may also have community-level and individual-level components. Without additional information, it is not possible to identify the impacts of unobserved individual-level determinants that are correlated among individuals within a community separately from the impacts of "independent," community-level unobservable variables and individual-level unobserved variables. As a consequence of this, we only consider two independent sources of unobserved or unmeasured variables. The first is a community-level unobserved factor that is independent across communities. We denote this type of factor by the random variable $\mathrm{E}_C(c)$; the realization of this random factor for community j is $e_C(c)$. The second type of unobserved factor is an individual-level random variable that is independent across individuals within a community. We denote these independent individual-level unobserved, random variables by $\mathrm{E}_I(i,c)$, and the realizations of these random variables by $e_I(i,c)$. Note that all of the observed variables in this study are assumed to be independent of all of the unobserved/unmeasured variables.

Using the notation defined above, we specify the general form of the data generating process (DGP) for our Monte Carlo experiments. In almost all instances, we examine models of the form:

$$Y(i,c) = 1\cdot X_C(c) + 1\cdot X_{IC}(i,c) + 1\cdot X_I(i,c) + \mathrm{E}_T(i,c)$$

where the composite error term $\mathrm{E}_T(i,c) = \mathrm{E}_C(c) + \mathrm{E}_I(i,c)$. Each of the observed explanatory variables is distributed as a normal random with mean zero and variance one, and the squared correlation of $X_c(c)$ *and* $X_{IC}(i,c)$ is set to 0.50. While the intercept in this DGP is 0, in all of the Monte Carlo experiments we do estimate an intercept. In our experiments we vary the R^2 across DGPs; we do this by choosing a variance for the composite error term E_T to yield the desired multiple correlation coefficient.[3]

[3]Note that $R^2 = 1 - \frac{Var(\mathrm{E}_T)}{Var(\mathrm{E}_T) + Var(X_C + X_{IC} + X_I)} = \frac{Var(X_C + X_{IC} + X_I)}{Var(\mathrm{E}_T) + Var(X_C + X_{IC} + X_I)}$.

A key component of this analysis examines how the performance of several estimators of the impacts of X_C, X_{IC}, and X_I vary as the influence of the community-level unobservables in the composite error increases. We vary the importance of the unobserved community-level characteristics by setting the fraction of the error variance due to the community-level unobserved determinants. In the multilevel modeling literature, this fraction is known as the intraclass correlation coefficient, and it is defined by

$$r = \frac{Var(\mathrm{E}_C)}{Var(\mathrm{E}_C) + Var(\mathrm{E}_I)} .$$

In the specifications of our data generating processes we assume that the components of the composite error term are distributed as mean zero, independent normal random variables, and we set the variances of $\mathrm{E}_C \; and \; \mathrm{E}_I$ to yield specified values of ρ.

Two special values of ρ are of particular interest. The first is when ρ equals 0. We obtain this by setting $\mathrm{E}_C(c)$ to 0 for all communities c (c=1,...,J). In this instance, the only source of error in the regression specification is from the independent, normally distributed, individual-level error terms. In this case OLS is the best linear unbiased estimator, and the simple standard errors of the estimators as reported by standard computer packages are correct. The second special value of ρ is 1.0, and we obtain this specification by setting the variance of $\mathrm{E}_I(i,c)$ to 0 for all individuals in all communities. In this instance, the error term is due entirely to community-level unobserved factors.

In nearly all experiments we estimate regression models of the form:

$$y(i,c) = b_0 + b_C x_C(c) + b_{IC} x_{IC}(i,c) + b_I x_I(i,c) \qquad (1)$$
$$+e_C(c) + e_I(i,c)$$

using observations on N_c individuals in each of J communities. The lower case Roman and Greek letters refer to particular realizations of the random variables, and we estimate the parameters β_0 (the intercept; true value 0), β_C (the impact of X_C, holding X_{IC} and X_I constant; true value 1), β_{IC} (the impact of X_{IC}, holding X_C and X_I constant; true value 1), and β_I (the impact of X_I, holding X_C and X_{IC} constant; true value 1).

For the most part we consider three estimation procedures for the parameter estimates and the associated standard errors of the estimates. The first is a simple, naive OLS estimation model that assumes the NJ observations are uncorrelated after controlling for the three explanatory variables when constructing the standard errors of the estimates[4]. This assumption is only correct when ρ equals 0, so the estimated standard errors from this procedure will be incorrect when ρ is different from 0. The second estimation procedure uses the OLS point estimates of the parameters, but it allows there to be arbitrary correlations among individual observations within each community when constructing the standard errors of the estimates. We do this by using Eicker-Huber-White standard error formulae.[5] The third estimation procedure is a random

[4]We do this, for example, by using the Stata command “regress y Xc Xic Xi” , where y is the outcome, Xc is the community level variable, Xic is the individual level variable correlated among individuals within the same community, and Xi is the independent individual level variable.

[5]We usually do this by using the Stata command “regress y Xc Xic Xi , cluster(c_id)” , where there three explanatory variables are as listed above and c_id is a variable that uniquely identifies each community. Note that these standard error estimators are consistent for arbitrary forms of heteroscedasticity. So even if the true DGP were a random parameter model, such as those used in more detailed multilevel models, these standard error estimators should provide

effects, maximum likelihood procedure that explicitly recognizes that observations within each community are equi-correlated with normally distributed errors[6]. Given that all error terms in this Monte Carlo analysis are normally distributed, have constant variances, and have a constant error structure across communities, no other unbiased estimator can provide more efficient estimates than this maximum likelihood procedure.

Monte Carlo Results

Preliminaries

For the first part of the Monte Carlo analysis we focus on sample sizes with approximately 20,000 individual-level observations. In the main body of the text we focus on specifications with 800 communities each containing between 1 and 50 individual observations, with a mean of 25 persons per community.[7] In the Appendix we present similar sets of results for 400 communities each containing exactly 50 individual-level observations. In a few instances a focus on the range of 1 to 50, or on exactly 50, individuals per community provides an incomplete view of the

---

unbiased hypothesis tests.

[6]We do this, for example, by using the Stata commands "`xtreg y` Xc Xic Xi`,i(c_id) mle`", where the three explanatory variables are as listed above and c_id is a variable that uniquely identifies each community.

[7]We used this range of individuals to represent roughly the distribution of the number of adult women per community in the Demographic and Health surveys. For each community in each replication of each data generating process we selected the number of individuals per community by taking draw from a truncated normal distribution with mean 25.5 and standard deviation 10 with the truncation points set at 1 and 50. We then took the integer portion of this truncated normal random variable as the choice of the number of individual level observations per community. This yields a mean number of individuals per community of 25 and a standard deviation of 9.5. Using this procedure, 91% of the time the number of individuals per community lies in the range [9,41].

impacts of the multilevel structure. In those instances we present numerical results from analytic formulae that indicate the impacts from varying the number of individuals per community (the number of level-one units per level-two unit) and the relationships among the explanatory variables. There is a general tendency for standard errors of estimates from the maximum likelihood estimator to be smaller than those from the OLS estimator, but this can vary by the type of variable being examined.

In the Monte Carlo experiments we focus on three values of the R^2 in the regression model: 0.10, 0.20, and 0.50. Not surprisingly, increases in the R^2 yield increases in the precision of all estimated parameters for all estimators, with corresponding decreases in standard errors. The primary conclusions of this analysis, however, are not affected by changes in the value of the R^2. We also examine twenty-one values for the intraclass correlation coefficient, ρ, for ρ equal to 0.00 to 1.00 by steps of 0.05.

For each specification of the data generating process we draw 1000 independent samples, each with 800 communities. Each community contains, on average, 25 individuals (standard deviation 9.5). We simulate community- and individual-level explanatory variables, community-level disturbances, and individual-level disturbances according to fixed, specific rules. For each of these 1000 replications of the DGP, we estimate the model specified in equation (1) by OLS and by a maximum likelihood procedure that allows for the hierarchical error structure. For the OLS estimates we calculate estimates of the standard errors of the point estimates by using standard, naive OLS formulae and by using the robust, Eicker-Huber-White formulae that adjust for the clustering within communities (i.e., possibly $\rho \neq 0$). For the maximum likelihood procedure, we use Stata's report of the square root of the diagonal elements of the inverse of Hessian matrix as the standard error estimator.[8]

Monte Carlo Evidence on the Unbiasedness of the Point Estimators of β_C, β_{IC}, and β_I

We treat the estimated coefficients and standard errors from each estimation approach for each of the 1000 independent samples as an independent draw from the distribution of the coefficient and standard error estimator for that estimation procedure. For example, if we look at the OLS estimator for a particular DGP, we can calculate the mean coefficient estimate for β_C as

$$\hat{b}_{C,OLS,DGP} = \frac{1}{1000}\sum_{e=1}^{1000}\hat{b}_{C,OLS,DGP,e}$$

where $\hat{b}_{C,OLS,DGP,e}$ is the OLS estimate of the coefficient on the community level observed variable for the e^{th} sample (replication) from a particular data generating process. If the OLS estimator is unbiased for this form of the DGP, then one would expect this mean coefficient estimate to be quite close to the true value specified for this DGP. If the mean coefficient were quite far from its known, "true" value as specified in the DGP, then one would suspect that the estimation procedure does not provide unbiased estimators for this form of the DGP.

[8] There is some empirical evidence that for values of ρ equal to 1.00 that Stata's standard error estimators for the maximum likelihood models exhibit some numerical instability. See, for example, Figures 5, 7, 8, and 9.

Figures 1A, 1B, and 1C provide graphical evidence on the unbiasedness of the OLS estimator and the maximum likelihood estimators for clustered data. Consider Figure 1A. This figure plots the mean estimates of the coefficient on the community-level variable, β_C , against the value of the intraclass correlation coefficient, ρ. Each of the three graphs in Figure 1A corresponds to a different level of the true R^2 in the regression model. The means of the OLS estimates $(\hat{b}_{C,OLS,DGP})$ are marked by a circle, and the means of the maximum likelihood estimates $(\hat{b}_{C,MLE,DGP})$ are marked by a plus sign(+). Figures 1B and 1C follow a similar format. These figures refer, respectively, to mean estimates of the coefficient on the individual-level variable that is correlated among community members (β_{IC}), and mean estimates of the coefficient on the independent individual-level variable (β_I). Recall that in each data generating process that we set the true level of each of these three coefficients, β_C , β_{IC} , and β_I , to 1.00.

Figure 1A reveals that both the OLS estimator and the maximum likelihood estimator appear to be unbiased. Additionally, the mean estimates from these two estimation procedures are nearly identical. Within each graph, higher levels of the intraclass correlation, ρ, appear to be associated with more variable mean estimates of β_C . This variability is due to the fact that the estimators are less precise at higher values of ρ (demonstrated in the next section) and to the design of our Monte Carlo experiments. Were one to use 100,000 or a million sample replications instead of only 1,000, this variability would be much less pronounced. At higher levels for the R^2 there is less variability in the mean estimates. This is as expected; each individual coefficient estimate is more precisely estimated with lower error variance. For each of the 63 data generating process represented in this Figure 1A (21 values of ρ and 3 R^2 values), the mean estimates are quite close to the true values. Similarly, Figures 1B and 1C reveal that the mean estimates of the impact of the correlated individual-level variable, $\beta_{IC,}$ and of the impact of the independent individual-level variable, β_I, are quite close to their true values of 1 for all values of ρ and the R^2. From these three figures one should conclude, in accordance to the predictions of least squares and statistical theory, that the OLS estimator and the maximum likelihood estimator are unbiased estimators of the effects of the community-level variables, correlated individual-level variables, and independent individual-level variables on the individual-level outcome, regardless of the fraction of the error variance due to the community-level unobserved factor or the R^2 value. Appendix Figures 1 provide similar evidence for the case of 400 communities with each with exactly 50 observations per community.

Monte Carlo Evidence on the Precision of the Point Estimators of β_C , β_{IC} , and β_I

One can easily calculate standard deviations of the coefficient estimates for an estimation procedure by using the sets of estimates obtained in the Monte Carlo experiments. These calculated standard derivations of the coefficient estimates should reflect the true sampling variability of the estimation procedure for particular specifications of the data generating process. In a similar fashion to the above analysis of the mean parameter estimates from the various estimation procedures, one could calculate, for example, the standard deviation of the OLS estimator for the DGP as

$$sd(\hat{b}_{C,OLS,DGP}) = \sqrt{\frac{1}{(1000-1)}\sum_{e=1}^{1000}(\hat{b}_{C,OLS,DGP,e} - \hat{b}_{C,OLS,DGP})^2}$$

This is an unbiased estimator of the standard deviation of the OLS estimator of β_C for this particular DGP from the 1000 replications in the Monte Carlo experiment. Provided that 1000 is a "large" number of replications, this estimate of the standard deviation should be close to the true standard deviation of the OLS parameter estimator[9]. Given that the evidence in the preceding section suggests that all of the point estimators are unbiased for the DGPs examined in this study, these calculated standard deviations should provide key evidence on the accuracy of the estimation procedures. For example, an estimator with a large standard deviation for a particular DGP would provide less accurate estimates than would an estimator that had a smaller standard deviation for the same DGP[10].

Figures 2A, 2B, and 2C provide Monte Carlo evidence on the accuracy of the OLS and maximum likelihood estimators of the three regression coefficients for three values of the R^2 with the intraclass correlation coefficient varying from 0 to 1. Figures 2 refer to samples of approximately 20,000 observations, where there are 800 communities each containing, on average, 25 level-one observations. Appendix Figures 2 contains similar information for the case

[9]In a real data set, the estimated standard error of the parameter estimate is meant to be an estimate of this true standard deviation of the estimator.

[10] It is important to recognize that these comparisons of the standard deviations as measured with the replications in the Monte Carlo experiments are meant to measure only the true sampling variability of the point estimators. These standard deviation estimates do <u>not</u> tell us directly whether particular <u>estimators of the standard error of the estimates</u> perform well. We examine that issue in the following section.

of 400 communities each containing 50 level-one units. The layout of these figures is the same as the layout for Figures 1A, 1B, and 1C.

Figure 2A indicates that the accuracies of the naive OLS estimator and the multi-level maximum likelihood estimator for the coefficient on the community-level variable are nearly identical for these specifications of the data generating processes. Especially for levels of the intraclass correlation coefficient less than 0.3, and regardless of the R-square value, there appears to be almost no efficiency gain from using the more exacting maximum likelihood approach instead of the simpler OLS approach. It is important to note that the standard deviations of the maximum likelihood estimator are slightly smaller than those of the OLS estimator, but for none of the 63 DGPs displayed in Figure 2A does the maximum likelihood estimator reduce the standard deviation by as much as 10%. Below we explore in more detail how the efficiency loss from using the less accurate OLS estimator varies as a function of the number of level-one observations, as the differences in accuracies become somewhat more pronounced when there are only a few level-one units per community.

Figure 2A reveals additional important information about the accuracy of both the OLS and maximum likelihood estimators of the impact of the community-level variable on the individual-level outcome. First, increases in the intraclass correlation coefficient cause a rapid decline in the accuracy of both estimators. Moving from independent observations (ρ=0.00) to an intraclass correlation of 0.25 causes the standard deviation of each estimator to approximately double. Stated differently, if there were a 25% intraclass correlation of the disturbances instead of independent disturbances, then all t-statistics for tests about the impact of the community-level variable would be about two times smaller than under independence; many fewer null hypotheses would be rejected. None of the above statements about the accuracies of the maximum likelihood and OLS estimators vary by the level of the R^2, though, as expected, the overall level of the standard deviation of the parameter estimator does fall for DGPs with higher levels of explanatory power.

A comparison of Figure 2A (800 communities, 1 to 50 individuals per community) to Appendix Figure 2A (400 communities, exactly 50 individuals per community) reveals that the relative performance of the OLS and maximum likelihood estimators remains nearly identical when there are more observations per community. The only change appears to be that the small advantage of the maximum likelihood estimator over the OLS estimator in standard deviation nearly disappears. In none of the 63 DGPs represented in Figure 2A does the standard deviation rise by more than 2% when one uses the relatively inefficient OLS estimator instead of the maximum likelihood estimator. Appendix Figure 2A also reveals that the efficiency loss from a higher intraclass correlation is somewhat more severe when there are more observations per community. Instead of the standard deviation increasing by a factor of two when ρ moves from 0.00 to 0.25, with the larger number of observations per community the standard deviations increase by about a factor of 2.5 to 3.[11]

Figure 2B provides comparable information about the coefficient on the individual-level variable that is correlated with the community-level explanatory variable, and Figure 2C provides the same type of information about the coefficient on the independent individual level explanatory

---

[11]Note that when $\rho=0.00$ that the standard deviations in Figures 2 and Appendix Figures 2 should be identical. With independence within communities, the precision of the estimators is not affected by whether the 20,000 observations come from 1 or 2 or 400 or 800 or 10,000 communities, provided that the explanatory variables follow the same distribution across communities.

variable. For both of these types of coefficients, there can be substantial improvements in the accuracy of the parameter estimates when one uses the multilevel maximum likelihood procedure instead of the simple OLS estimator. The proportionate increases in precision from using maximum likelihood for estimating these two coefficients appear quite similar in the Monte Carlo experiments.

The analytic variance formulae presented in the Appendix indicate that the asymptotic variances for the estimators of the two individual-level variable effects differ only because, after controlling for the community-level variable, there is less independent variation in the correlated individual-level variable than there is in the independent individual-level variable. Analytically, each of the standard deviations presented in Figure 2C are smaller than the corresponding standard deviations in Figure 2B by a factor of $\sqrt{1-t^{2}}$, where ^{2} is the squared correlation of the community-level variable and the correlated individual-level variable (equal to .5 in the Monte Carlo experiments). Given this exact correspondence of the standard deviations of the estimators for these two coefficients for both estimation procedures, we only focus on the coefficient of the correlated individual-level variable in the following discussion.

A comparison of Figures 2A and 2B reveals two important differences in the performance of the estimators of the coefficients for the community-level variable and the estimators of the coefficients for the individual-level variables. First, the estimators of the impact of the community-level variable becomes less precise with increases in the intraclass correlation, while the precision of the OLS estimator for the impact of the individual-level variable does not change as the intraclass correlation increases. Second, there are important efficiency gains from using maximum likelihood estimators instead of OLS estimators when estimating the impacts of variables that vary within the community. Appendix Figures 2A and 2B provide similar information for the case of 400 communities with exactly 50 observations per community.[12]

There is a simple explanation for the first difference in the performance of the estimators. As one increases the level of correlation of the disturbances within a community, there is less new information provided by each observation within the community. The accuracy of the correlation of the community-level variable and the disturbance for each observation, then, is directly impacted by this decrease in information. Individual-level variables, on the other hand, do have unique, independent variation within a community. Heuristically, for the individual-level variables, each additional observation in a "new" community provides the same additional information as adding an individual to an already existing community.

One can make this latter point a bit more formally. By definition, the OLS estimator for each coefficient depends only on the variation in the explanatory variable that is linearly independent of the other explanatory variables in the regression equation and its interaction with the composite error term. Increases in the intraclass correlation decrease the precision of the estimated correlation of the disturbance with the community-level variable, and this leads to the OLS estimator of the impact of the community variable becoming less precise as the intraclass correlation rises. Increases in the intraclass correlation, however, do not affect the precision of

[12]The only discernable, substantive difference between 400 communities and 800 communities results from the fact that the 800 communities provide twice as many independent observations on communities, and and so yields increases in the accuracy of the parameter estimates. Almost no other comparison varies substantively along these community size and number of individuals per community dimensions, so we do not report any additional results for the 400 communities with each containing 50 observations.

the correlation of the disturbance and the linearly independent variation in the individual-level explanatory variable. This happens because the individual-level variable has unique, independent variation within a community, and products of this independent variation with the unobserved community-level factor are independent across observations regardless of the level of the intraclass correlation. Consequently, the precision of the OLS estimator of the impact of the individual-level variables is unaffected by changes in ρ.[13]

The failure of the maximum likelihood estimator to improve the precision of the estimates of the impact of the community-level variable can best be understood, by analogy, by considering the efficiency gains from seemingly unrelated regression (SUR) estimators. When there is error correlation across observations, the SUR estimator can only yield increased precision over the OLS estimator if the explanatory variables are not identical across the observations with the correlated error terms. This happens because the SUR estimator exploits the fact that explanatory variables for each particular observation should be uncorrelated with the error terms for all observations that are correlated with that particular observation's error term. If the explanatory variables are identical across these observations, then there are no new correlations of explanatory variables and error terms within clusters that were not used to define the OLS estimator. For the DGPs considered here the individual-level variables do differ across observations within clusters, so there are some efficiency gains for the estimators of the coefficients on these individual-level variables. But almost no new information is provided by imposing the restriction that the

[13] If, however, there is some correlation within communities of the individual level variables that is not captured entirely by the included community level explanatory variables, then the precision of the OLS estimator of the impact of the individual level variables will decline with increases in the intraclass correlation coefficient.

community-level variable for an observation within a cluster is uncorrelated with all the error terms for all observations within the same cluster; that information had already been used with the OLS estimator because the community-level variable is constant within the cluster.

Impacts on Precision from Varying the Number of Individuals Per Community

These statements about the relative efficiencies of the estimators, as measured through their standard deviations, do depend on the number of individuals per community in a somewhat complex manner. To explore this in more detail, we use analytic expressions for the asymptotic variance covariance estimators for the point estimators. We assume that there are an identical, finite number of observations per community. With normally distributed residuals, the maximum likelihood estimator converges to the (GLS) Generalized Least Squares estimator using the true covariance matrix of the residuals within a community when there are a large number of communities. Because of this we can use an analytic expression for the asymptotic covariance matrix of this GLS estimator to proxy for the large sample covariance matrix of the maximum likelihood estimators. We also calculate the true asymptotic covariance matrix for the OLS point estimators. This covariance matrix for the OLS estimators accounts for the fact that there are multilevel disturbances even though this was ignored in the definition of the estimator. All of these analytic expressions allow there to be an arbitrary level of intraclass correlation, an arbitrary level of correlation between the community-level variable and the correlated individual-level variable, and for an arbitrary number of individuals per community. All these three factors are important determinants of the two covariance matrices. A comparison of these asymptotic covariance matrices tells us precisely the efficiency gains from using maximum likelihood instead of OLS estimators. The derivations of the asymptotic covariance matrices are in the Appendix.[14]

Figures 3A and 3B graph how the standard deviations of the OLS estimators of the impacts of the three variables, relative to the standard deviations of the maximum likelihood estimators, vary by the number of observations per community and the level of correlation of the community variable and the individual-level variable at two particular values of the intraclass correlation.[15] Figure 3A examines the case when the intraclass correlation, ρ, is 0.25, and Figure 3B examines the case for ρ=0.75. Each of the four graphs within these figures refers to a different value of the correlation between the community-level explanatory variable and the correlated individual-level explanatory variable (0.00, 0.25, 0.50, and 0.95).[16] The horizontal axis measures the number of observations per community. The vertical axis measures the standard deviation of the OLS estimator as a fraction of the standard deviation of the efficient maximum likelihood estimator. This provides a measure of how much efficiency loss one can expect by

[14]Note that we arbitrarily set the overall error variance to 1.0 in these covariance matrixes. Since we only examine ratios of standard deviations, this is completely inconsequential. The Appendix also contains the formula for the simple, naive OLS covariance estimator that ignores the multilevel structure of the residuals.

[15]We evaluate the analytic formulae for the variances of the two estimators of each regression coefficient, calculate their ratio and take the square root. This provides a ratio of the standard deviations of the estimators. A value of 1.10, for example, would mean that the OLS estimator would have a standard error of estimate 10 percent higher than the maximum likelihood estimator of the coefficient for the same DGP; heuristically, t-statistics would tend to be about 10% smaller for OLS estimator than they would be for maximum likelihood estimator.

[16]Recall that in all of the previous experiments we kept the level of correlation of these two regressors fixed at 0.50.

using the less precise OLS estimator instead of the maximum likelihood estimator. The relative efficiencies for the estimators of the impacts of the two individual-level variables (indicated by the diamonds and plus signs) are identical analytically, and they do not depend on the degree of correlation between the individual-level variables and the community-level variable.

At the moderate level of the intraclass correlation in Figure 3A, ρ=0.25, there would be little efficiency gain from using the maximum likelihood estimator instead of the OLS estimator. The maximum gain in t-statistics as implied by these relative standard errors, for example, would be less than 15%, and such gains could only be attained for the estimators of the impacts of the individual-level variables. When there is no correlation among the community-level variables and the individual-level variables, there are zero efficiency gains from using maximum likelihood in the estimation of the impact of the community level. Only when the correlation of the community-level variable and the individual-level variable is quite high is there any discernable efficiency gain for the estimator of the impact of the community-level variable from using maximum likelihood estimation.

The efficiency gain for the estimation of the impact of the community-level variable initially increases as one adds more observations per community, but then it falls. But with ρ=0.25, even when the correlation of the regressors is as high as 0.95 the standard error improves by using maximum likelihood by less than 10 percent. For the individual-level variables, the efficiency gains do increase continually as one adds additional observations per community, but there are only trivial efficiency gains after having 25 or 30 observations per community. At this level of intraclass correlation, none of the efficiency gains from using maximum likelihood estimation instead of OLS estimation is substantial.

Figure 3B examines the case where there is a high level of intraclass correlation, ρ=0.75. There are potentially sizable gains in precision by using maximum likelihood when estimating the impacts of the individual-level variables in these circumstances. By using maximum likelihood instead of OLS, one could reduce standard errors of the estimators of the individual-level variables by about a factor of two. As above, however, there is little efficiency gain from using maximum likelihood to estimate the impact of the community-level variable, unless the correlation of the community-level variable and the individual-level variable is quite high. But even when there can be substantial efficiency gains in estimating the community-level variable impact by maximum likelihood, the gains diminish rapidly with increases in the number of observations per community.

The interaction of the number of observations per community, the intraclass correlation, and the correlation of the community-level regressor with the individual-level regressor appears to be the key determinant of efficiency gains from maximum likelihood estimation when estimating the impact of the community-level covariate. In Figures 4A and 4B we examine this relationship in finer detail along the dimension of the correlation of the community-level variable and the correlated individual-level variable. As in Figures 3, the top panel in Table 4 is for ρ=0.25, and the lower panel is for ρ=0.75. The graphs in each figure are for different numbers of individuals per community (NIPC=2, 5, 25, and 50). The horizontal axis measures the level of correlation between the explanatory variables( ). In Figure 4 we only examine the relative efficiency for the estimators of the impact of the community-level variable.

For the most part, there appear to be almost no efficiency gains from using maximum likelihood estimators instead of OLS estimators for values of the regressor correlation being less than 0.50. For the moderate level of the intraclass correlation, 0.25, there never are efficiency gains over 15 percent for all values of the regressor correlation at 0.99 or lower. When the intraclass correlation is high, ρ=0.75, there can be some substantial gains in efficiency, with the larger gains happening when there are several individuals per community. Note, however, that these gains are quite small unless the regressor correlation is well over 0.50.

Summary of the Accuracy of Estimators with Multilevel Errors

The information summarized in Figures 1 through 4 and the Appendix provide key information on the importance of controlling for error correlations due to a researcher having hierarchical data. First, coefficient estimates are not biased if one ignores the multilevel error structure and uses a standard OLS model to estimate the impacts of community-level and individual-level covariates on an outcome. Second, as is well known in the survey design literature, there can be important losses in efficiency when data are clustered and the intraclass correlation increases (Kish, 1965; Kalton, 1983). These efficiency losses due to increased within cluster error correlations, however, only take place for estimators of the impact of the community-level variable. In fact, a clustered design can yield efficiency increases for the estimated impacts of individual-level variables if one uses a maximum likelihood procedure that recognizes the intraclass error correlations. For many analyses of programmatic impacts, however, the effects of interest are not those of the individual-level variables. Obtaining efficiency gains for the impacts of the individual-level variables is at most of secondary importance, so this efficiency gain by itself does not provide a compelling reason to use maximum likelihood estimators instead of OLS estimators. These efficiency gains are substantial for the estimated impact of the community-level variable (decreases in standard errors by more than 15 percent from using maximum likelihood instead of OLS) only when both the intraclass error correlation and the correlation of the regressors are high.

Third, it is the correlation of the community-level variable with the individual-level variable that provides all of the efficiency gain in the estimation of the community-level variable impact from using maximum likelihood estimation instead of OLS. With moderate sized regressor correlations there are at best small efficiency gains from using maximum likelihood to estimate the impact of the community-level variable instead of OLS. For both of these correlations being 0.50, for example, the decrease in the standard deviation is less than 2.5 percent from using the maximum likelihood estimator. For ρ=0.50, this measure of the efficiency gain only exceeds 15 percent if the community-level and the individual-level regressor have a correlation higher than 0.90; and for a regressor correlation of 0.50, the standard errors decline by less than 8 percent for all values of the intraclass correlation lower than 0.99. Additionally, the efficiency gains in the estimation of the impact of the community-level covariate decline after there are more than 5 or 10 observations per community.

The overall picture that emerges from these Monte Carlo experiments and the examination of the asymptotic covariance matrices is that there appears to be little efficiency loss from using OLS instead of maximum likelihood to estimate the impact of a community-level variable on an individual-level outcome. The only substantial gains come when both the intraclass correlation and the correlation of the community-level regressor with the individual-level regressor are exceptionally high. We suspect it is unlikely in most situations for both of these correlations to be high, so the loss in efficiency from using Ordinary Least Squares should usually be quite small.

Evaluation of Estimators for the Standard Errors of the Estimates

In this section we examine the performance of estimators of the standard errors of the point estimators for the three coefficients. When examining estimators of standard errors, it is important to recognize that one should not examine whether the mean of the standard error estimator equals the true standard deviation of the estimator[17]. Instead, one should assess whether hypothesis tests that use the standard error estimator yield accurate probabilities under the null hypothesis. In particular, a standard error estimator would be considered accurate if hypothesis tests that use this estimator reject a true null hypothesis with a frequency given by the specified size of a test. If one tests at the 5% level, for example, then one should reject correct null hypotheses 5% of the time. Otherwise the standard error estimator does not allow one to carry out precise tests.

A standard, simple hypothesis test is of the form: $H_0{:}\, b = b_0 \; vs \; H_1{:}\, b \neq b_0$. One typically undertakes a hypothesis test of this form by using a two-tailed test under the assumption that the estimator of β follows an approximate Student T or normal distribution. To carry out such a test, one sets a size of the test, ; this is the specified probability of rejecting the null hypothesis when

[17]For example, the standard OLS estimator of the covariance matrix for the point estimators provides unbiased estimators of the variances of the parameter estimators under the assumptions of the classical regression model, namely, the diagonal elements of , $\hat{s}^2 \; (X'X)^{-1}$. The estimators of the standard errors given by the square roots of the diagonal elements of the covariance matrix, however, must be biased estimators of the standard errors of the estimators (by Jensen's Inequality, since the square root is not a linear transformation of the unbiased variance estimator, the standard error must be a downwardly biased estimator of the true standard deviation).

the null hypothesis is actually true.[18] Our evaluation of the accuracy of the estimator for the standard error of the estimate is an assessment of how closely the frequency of rejecting a true null hypothesis in the Monte Carlo experiments matches the specified size . If we specify   = 0.05, then we would want to have the null hypothesis that $\beta=\beta_0$ to be rejected five percent of the time when β actually does equal the value β_0 .

In the Monte Carlo experiments we know exactly the true parameter value (all β's equal to 1), so we can examine how frequently a true null hypothesis is rejected when we use various standard error estimators for the different point estimation procedures. We examine the size of the tests for three configurations of the test. If the intraclass correlation is zero, all testing configurations should provide close to identical results.

Two configurations use the ordinary least squares point estimators for the parameter estimates. The first of these uses the standard error of the estimate as reported by the simple ordinary least squares procedure to evaluate the hypothesis test. This procedure corresponds to using a standard OLS procedure and using the "default" estimators of the standard errors that assume uncorrelated, homoscedastic disturbances. This will usually provide biased tests when the intraclass correlation is non-zero for the DGPs examined here.

The second testing configuration uses a robust standard error estimator that accounts for the fact that there could be arbitrary correlations of disturbances within communities along with

---

[18]We carry out this hypothesis test by examining whether the shortest 100*(1-)% confidence interval for the estimator contains the true parameter value. In particular, we construct the confidence interval, assuming that the parameter estimates are approximately normally distributed, as $(\hat{b} - z_{1-a/2}\,\hat{se}(\hat{b}),\ \hat{b} + z_{1-a/2}\,\hat{se}(\hat{b}))$ and reject the null hypothesis if this region does not contain the value specified under the null hypothesis. The term $z_{1-\ /2}$ is the 100(1- /2) percentile point of the standard normal distribution.

the simple OLS parameter point estimator (Eicker, 1963, 1967; Huber, 1967; White, 1980). These Eicker-Huber-White standard error estimators require that one specify a hierarchical level such that there is independence of disturbances across groups at all higher levels. For example, if individual-level residuals were correlated within families but not across families living in the same community, one would specify the family as the highest level where there is correlation.[19] These Eicker-Huber-White standard error estimators also allow for arbitrary forms of heteroscedasticty, so they would provide appropriate standard errors in the presence of random coefficient models.

The third testing configuration we examine uses the maximum likelihood estimator. We use the point estimators and the standard error estimators from the maximum likelihood procedure to carry out hypothesis tests. For each of these three testing configurations, we examine whether the standard error estimators used with the point estimators provide tests of the correct size.

Figure 5 provides evidence on the probability that each of the three different testing procedures incorrectly rejects a true null hypothesis.[20] Since we carry out 1000 Monte Carlo replications, the vertical axis measures the fraction of times out of the 1000 replications that the hypothesis test rejects a true null hypothesis for a particular testing procedure. The horizontal axis

[19]If there were some community level unobserved variables that influenced the individual outcomes, then specifying the family as the highest level with correlation would be incorrect. In a later section we explore the consequences of specifying this level as either too high or too low.

[20]The tests in all of the graphs in Figure 5 use data for the case where there are 800 communities with the above specified distribution for individuals per community (mean 8, standard deviation 9.5). These evaluations of the size of the tests do not depend upon the R^2 value, as changes in the error correlation increase the variability of the point estimate by precisely the same proportion as the error variance changes. For the 1000 replications used in these graphs, we select arbitrarily the $R^2=0.10$.

measure the level of the intraclass correlation coefficient. The left-hand set of graphs presents tests where the desired size of the test is five percent (0.05), and the right hand set of graphs contains tests where the desired size of the test is ten percent (0.10). The top row of graphs corresponds to tests for the coefficient on the community-level variable, the second row of graphs to tests for the coefficient on the individual-level variable that is correlated across individuals within each community, and the third row of graphs to tests for the coefficient on the independent individual-level variable.

The top, leftmost graph in Figure 5 examines the performance of the three testing configurations for testing the null hypothesis that the coefficient on the community-level variable equals its true value as specified in the data generation procedure (against the null hypothesis that the coefficient does not equal its true value). When the intraclass correlation coefficient (ρ) equals 0.00, all three of the testing procedures yield approximately the correct size of 0.05. As the intraclass correlation rises, the procedure using the ordinary least squares point estimate with the simple OLS standard error estimate (labeled olstest, with circles) has an empirical probability of false rejection that greatly exceeds the specified 5% size. For all intraclass correlations above 0.10, the empirical size of this testing procedure exceeds 20% when one specifies a probability of false rejection of only 5%. With fewer communities and the same number of total individual-level observations, the empirical size from this approach can be much greater than 50% for a specified size of 5%.

From the same graph in Figure 5, tests that use the same OLS point estimate as above for coefficient on the community-level variable but with the standard error adjusted to correct for arbitrary forms of correlation within communities (labeled osthtest, with triangles) yield approximately the correct size for all values of the intraclass correlation coefficient. Similarly the tests based upon the maximum likelihood point estimates and the corresponding maximum likelihood estimates of the standard errors of the estimates appear to have approximately the correct size. The top right-hand graph provides similar evidence for the case where the requested size is increased to 10%. Little of substance changes with this increase in requested size. In summary for the community-level variable, relying on the OLS point estimates and simple, default standard error estimates results in hypothesis tests that too frequently reject true null hypotheses. The simple OLS standard error estimates are, in a sense, too small. This propensity to reject null hypotheses much too frequently can be fixed with either of the other approaches. One can use the same OLS point estimates in conjunction with Eicker-Huber-White standard errors estimators to accommodate possible residual correlations within communities. Or, one can use maximum likelihood point and standard error estimators that correctly specifies the form of the within community disturbance correlation.

The second row of graphs in Figure 5 examines sizes of tests for the coefficient on the individual-level variable that is correlated across individuals within a community. Each of the three testing procedures yields tests that have approximately the correct empirical size for all values of the intraclass correlation and for both requested size levels, 0.05 and 0.10. The fact that the OLS point estimates used in conjunction with the naive OLS standard error estimate that do not yield an incorrect empirical sizes is surprising. But if one examines the true standard deviation of the OLS estimator for this correlated individual-level variable and compares it to the corresponding element of the $(X'X)^{-1}$ matrix, both presented in the Appendix, it is clear that the naive standard error reported by OLS is correct for this variable. The third row of graphs in Figure 5 reveals similar behavior for the empirical sizes associated with test of the coefficient on the independent individual-level variable. Each of the three approaches for testing hypotheses about individual-level coefficients has the correct empirical size.

It is important to note that the empirical size for the tests of the coefficients on the individual-level variables will, in general, be incorrect if one uses the ordinary least squares estimate in conjunction with the simple standard error of the estimate reported by the OLS procedure. In the Monte Carlo experiments examined here, all correlations of the individual-level variables were due to the observed, and controlled for, community-level variable. If there were other, independent community-level factors that gave rise to correlations of individual-level variables, then the empirical size from the OLS point estimates with simple OLS standard errors of the estimates would be much larger than the requested size. The two other testing approaches are not affected by the source of the correlation of individual-level covariates within communities. The OLS point estimator with a standard error that adjusts for arbitrary community correlations of disturbances will provide correct sized tests, as will the maximum likelihood point and standard error estimators.

These examinations of the empirical sizes of the tests reveal one important point, namely that tests using naive standard error estimators based upon unfounded assumptions about uncorrelated disturbances can indicate much more precision of parameter estimates than is actually warranted. This is clearly an important issue for the estimated impact of the community-level variable. It is also an important factor to consider when the observed individual-level variables are correlated within communities in ways that are not captured completely by the observed community-level variable. A key point to note is that the maximum likelihood procedures and the OLS point estimates with standard errors estimators that permit arbitrary correlation within communities do provide tests whose empirical sizes appear to be correct.

Power Comparisons of OLS and Maximum Likelihood Estimators

Given that there are estimators and testing procedures that appear to have the correct size for the forms of multilevel models that we have examined, we can now examine the ability of these procedures to reject null hypotheses that are incorrect. Holding the size of the test constant, one would prefer to have estimators and procedures that reject false null hypotheses more frequently. The ability of a test taking the form $H_0\!:b = b_0 \ vs \ H_1\!:b \neq b_0$ to reject the null hypothesis when the alternative is in fact correct depends crucially on the true value of the parameter. If the true value is quite close to the value specified under the null, then the probability of rejection is quite close to the specified size of the test (e.g., only five or ten percent), while if the true parameter value differs dramatically the probability of rejection should be close to 1.0. A graph displaying the probability of rejecting $H_0:\beta=\beta_p$ versus $H_1:\beta\neq\beta_p$ for a possible set of values β_p when the true parameter $\beta=\beta_0$ is one way of displaying the power of the test. We assess the power of each test empirically by using the estimates and standard errors from the Monte Carlo experiments. For each null hypothesis examined, we present the fraction of times (out of 1000 replications) that the testing procedure would reject the null hypothesis.

Figures 6 contain graphs of the power functions corresponding to tests for each of the three regression coefficients of the form $H_0\!: b = b_p \ vs \ H_1\!: b \neq b_p$. The definition of the power function displayed here is the probability of rejecting the null hypothesis $\beta=\beta_p$ (against the alternative $\beta\neq\beta_p$) as a function of the value of β_p when the true parameter value is 1.0.[21] Figure 6A examines power functions for hypotheses about the coefficient on the community-level parameter; each of the four graphs within Figure 6A corresponds to a different level of the intraclass correlation. Figure 6B contains similar information about tests for the coefficient on the within community, correlated individual-level variable, and Figure 6C displays the power functions for tests about the coefficient on the independent individual-level variable. We set the size of all these tests to 0.05. Here we only examine those approaches with correct size for arbitrary levels of the intraclass correlation, so the power function evaluated at β_p=1.0 equals 0.05 for all testing approaches displayed here.

For the community-level coefficient only two testing approaches had the correct size: OLS point estimates with standard errors adjusted for possible clustering of disturbances within communities and maximum likelihood point estimates and standard errors. These are displayed in Figure 6A as “olshtest” and “mletest.” When the intraclass correlation is 0.10, one would reject

---

[21]This definition differs from the usual definition of a power function. For the more standard definition of the power function, one tests an identical hypothesis (e.g. H_0: =2 vs.H_a: $\neq$2) and graphs the probability of rejection as a function of a varying true value of the parameter. Here, we graph the probability of rejecting a varying null hypothesis, when the true parameter value 1.0, as a function of hypothesized values specified in the null and alternative hypotheses.

the null hypothesis that β_c=0.75 (or β_c=1.25)[22] about 76% of the time with OLS and 79% of the time with the maximum likelihood procedure when the true value of β_c=1. As the intraclass correlation rises to 0.25, the power to reject H_0: β_c=0.75 (or β_c=1.25) when the true value is 1 falls to 50% for the OLS-based procedure and 55% for the maximum likelihood procedure; when the intraclass correlation is a high 0.75, the power for the same test is only 25% for the OLS based test and 27% for the maximum likelihood based test. In all cases examined here, the largest discrepancy in size between the two testing procedures is only eight percentage points. In over half of the tests displayed in the top panel of Figure 6, the probability of rejection using the maximum likelihood estimator is less than one percentage point larger than the probability of rejection from using the OLS point estimate with the Eicker-Huber-White standard error.

Overall, one would conclude that there is little difference between the performance of tests based on the OLS point estimates of the coefficient on the community-level variable (and with the standard error adjustment for arbitrary within community correlation) and the tests based on the maximum likelihood estimates. This should not be surprising; the comparisons of the asymptotic standard deviations of these two estimators discussed above indicates that there would only be sizable differences if the intraclass correlation were exceptionally high with only a few observations per community.

Figure 6B contains similar information on the power of tests on the coefficient of the individual-level variable that is correlated at the community level. Recall for this coefficient estimate that the simple, naive OLS standard errors that do not recognize the correlation of the

[22]The alternative hypothesis for all power and size tests discussed in this study is the complement of the null hypothesis under examination.

disturbances within a community provide tests that yield the correct size. These standard error estimates are nearly identical to those provided by allowing for arbitrary forms of correlation within communities, and we focus on the adjusted standard errors in the following discussion. Note that the horizontal scale in Figure 6B is much smaller than the horizontal scale in Figure 6A; we use this smaller scale because the estimators of this coefficient are much more precisely estimated.

At low levels of intraclass correlation the differences in the power between tests using the OLS point estimates and tests using the maximum likelihood point estimates are small for tests about the impacts of the individual-level variables. Consider for example, an intraclass correlation of 0.10 and one tests $H_0\!: b = 0.90$ *vs* $H_1\!: b \neq 0.90$ when the true parameter value is 1.0. A 5% test based on the OLS point estimate and corresponding standard error would reject this false null hypothesis about 37% of the time, while the maximum likelihood estimates would reject the null hypothesis about 41% of the time. Recall that the standard deviation of the OLS estimator does not vary with the level of the intraclass correlation; the power function for the OLS estimator should not change as one varies the intraclass correlation. The maximum likelihood estimator, however, makes efficient use of the fact that disturbances are correlated and that these explanatory variables differ across observations within the community.[23] Consequently, the power of tests on the individual-level variable impacts based on the maximum likelihood estimates increases substantially as the intraclass correlation (ρ) rises. For the above hypothesis test, when

[23]This is precisely a seemingly unrelated regression framework, and increased power from the maximum likelihood approach reflects the precision gain from using seemingly unrelated regression.

the true value of the coefficient is 1.0, the power rises to 47% at ρ=0.25, then to 61% for ρ=0.50, and to over 90% at ρ=0.75. Unlike the case for the estimated impact of the community-level variable, there are clear gains in the precision of the estimates of the impact of the correlated individual-level variable that one can obtain by using maximum likelihood procedures instead of OLS point estimates.

The power functions for the estimators of the coefficient on the independent individual-level variable, as displayed in Figure 6C, are qualitatively the same as those depicted for the correlated individual-level variable coefficient in Figure 6B. In fact, the only substantive difference between these two sets of figures is due to the fact that the estimators of the coefficient on the independent-level variable have smaller standard errors. This happens because these explanatory variables are not correlated with any of the other explanatory variables. For low levels of ρ there are at most modest increases in power due to a researcher using the maximum likelihood procedures instead of OLS, while there can be fairly large increases in power if there is a high level of correlation of the unobserved factors within communities.

We also examined a few situations where an individual-level regressor is correlated across members within each community but the correlation is not due entirely to the observed community-level variable. The value of using the maximum likelihood estimators instead of the OLS estimators for the estimation of individual-level covariate effects can increase substantially in these instances. It is important to note that any large efficiency gains only apply to the estimation of the impacts of individual-level variables. The gain in precision for the estimates of community-level characteristics by using the more complex maximum likelihood approach is typically quite small.

Spurious Intraclass Correlation

Many forms of model misspecification can result in estimates suggesting falsely that there is an important multilevel error structure when the true error correlation is actually zero. An erroneously excluded community-level variable, for example, can give rise to spurious evidence about the importance of a multilevel error structure. In this case, if the excluded variable is at all linearly related to any of the included explanatory variables within possible clusters, then the parameter estimators will be biased. The evidence of a non-zero intraclass correlation for this instance could be an indication of an incorrectly specified model with no multilevel structure, while a naive interpretation of the evidence would conclude that controlling for the multilevel error structure is important for one to obtain accurate estimates.

Variations across communities in the distribution of any of the individual-level characteristics can also lead to spurious evidence of a multilevel error structure in these situations. It might be the case, for example, that people tend to live in communities that have educational backgrounds similar to their own. If one specifies an incorrect functional form for how these individual characteristics (e.g., individual education) influence the outcome, then one will typically estimate a large intraclass correlation even if the true error correlation is actually zero.[24]

Model misspecification problems, of course, are not unique to multilevel analyses. They pervade all empirical analyses. What is somewhat unique to the multilevel model framework is that an incorrect specification of the regression function can easily provide evidence that one could interpret as indicating the presence of a complex, hierarchical error structure, when the only

[24]For example, this could be the case when the true relationship is quadratic but the researcher only allows for a linear relationship.

problem is a misspecified regression function. A researcher focusing on the multilevel structure of the residuals could easily fail to recognize a significant specification error for how the observed covariates impact the outcome of interest after uncovering what appears to be a significant level of intraclass error correlation. This can lead to quite biased estimates and interpretations.

Consider the model

$$Y(i,j) = b_C \cdot X_C(j) + b_{C2} \cdot [X_C(j)]^2 + b_{IC} X_{IC}(i,j) + b_I \cdot X_I(i,j) + \mathrm{E}(i,j)$$

where the error terms are uncorrelated across individuals within the same community (i.e., ρ=0.00). This regression specification differs from those examined above only by the inclusion of the square of the community-level characteristic. For this specification of the model, an ordinary least squares regression of Y on X_C, the square of X_C, X_{IC}, and X_I would yield the best linear unbiased estimated of the coefficients.

One might be interested in how the impact of X_C varies at different levels of the explanatory variable, and this is given by the first derivative $b_c + 2 \cdot b_{c2} \cdot X_C$. An alternative, scalar measure of the impact of X_C is the average derivative given by $b_C + 2 \cdot b_{C2} \cdot \bar{X}_c$ where $\bar{X}_C$ is the mean value of the community-level variable. We focus on this scalar measure as researchers often assume (incorrectly) that simple models excluding the higher order terms do capture the average impact of variables.

In the data generating process we set $\beta_C=\beta_{IC}=\beta_I=1$ and $\beta_{C2}= -0.50$. We also set the intraclass correlation to 0.00. All other aspects of the data generating process are as above, except we use a skewed distribution for the X_C variables. We also impose that the X_C variables are fixed across replications of the Monte Carlo experiments.[25] This eliminates a source of variation in the calculation of the average impact of the community-level covariate. Given this specification the true average impact of the community-level variable is 1.028.[26]

Based on 1000 replications of the DGP, a simple OLS model that fails to include the square of the community-level variable yields a mean estimate of the average impact of 0.085. The standard deviation of these 1000 estimates of the average impact is 0.010. All 1000 naive estimates of the standard error of the average impact all fall in the range (0.0108,0.0112), indicating that the standard error estimators reflect fairly well the variability of the estimator. This simple model, however, does underestimate the true average impact of the community-level variable by a factor of 12, but one would conclude that the effect is significantly different from zero.

Using a maximum likelihood procedure to allow for a hierarchical error structure with this misspecified model, from the 1000 replications of the data generating process we find that the mean of the average impact of the community-level variable remains at 0.085, with only a slightly smaller standard deviation (0.009). The 1000 estimates of the standard error of the average effect, however, all fall in the range (0.0668,0.0723), indicating that the maximum likelihood

[25] We have these community level variables follow approximately a Chi-Square distribution with 1 degree of freedom. Precisely, for each community j=1,2,...,J, we set the value of the community level variable to be the j/(J+1) percentile point of the Chi-Square distribution. We normalize this to have approximately mean zero and variance 1 by subtracting the mean of a Chi Square random variable (mean=#degrees of freedom) and divide this by the standard deviation of a Chi-Square random variable (variance=2 times the #degrees of freedom).

[26] Following the above formula for the average derivative, 1.028=1+2(-0.5)(-0.028), where -0.028 is the mean of the community level variable as specified above.

multilevel model severely overestimates the true sampling variability of the estimator. The largest t-statistic out of the 1000 tests of the null hypothesis that the true average effect is 0 is 1.68; in not one of the 1000 cases would one have concluded that the community level variable had a significant effect at conventional significance levels. One finds, however, that allowing for a multilevel error structure is important. All 1000 of the estimated intraclass correlations fall in the range (0.259,0.293), even though the true intraclass correlation is 0. The multilevel model yields quite biased estimates and inferences when one does not use the correct specification for the regression function.

It should not be surprising that the multilevel model fails to perform well in this instance. The primary benefit of the multilevel model is to obtain correct standard errors of the estimates and to obtain more precise parameter estimates. The maximum likelihood, multilevel model approach cannot fix functional form problems, and in this instance the approach actually appears to provide more incorrect interpretations than would come from a simple OLS estimation. Functional form problems appear to be more important issues to address than problems arising from the correlation of disturbances within communities.

A simple way to allow for flexible forms is to use polynomials of the explanatory variables and interactions among these polynomials[27]. Consider the following example. As above, the true model in the DGP is

$$Y(i,c) = b_C \cdot X_C(c) + b_{C2} \cdot [X_C(c)]^2 + b_{IC} X_{IC}(i,c) + b_I \cdot X_I(i,c) + \mathrm{E}(i,c)\,,$$

[27]Many other approaches could be used, such as spline approximations or neural networks, but often these are more difficult to implement in standard statistical packages.

but we assume we do not know that this is the correct functional form. Instead, we enter a fully interacted cubic polynomial in the three explanatory variables. While the true model would include only five regression parameters to estimate (including the intercept), this fully interacted polynomial introduces a total of twenty parameters that need to be estimated. The first derivative of this polynomial includes ten of these estimated parameters, compared to only two estimated parameters needed if one knew the true form of the regression model (i.e., β_C and β_{C2}). We impose that the community characteristics are fixed, that the individual-level characteristics vary across replications of the DGP, and that there is no intraclass correlation of error terms. We set the R^2 equal to 0.20 and use 100 communities each with 5 level-one units, for a total of only 500 observations.

Estimating the true model yields an average estimate of the impact of the community covariate of 1.027 across the 1000 Monte Carlo replications, with a standard deviation of 0.087. Recall that the true average effect is 1.028. When we estimate the fully interacted cubic polynomial model, the estimates have an average value of 1.020 with a standard deviation of 0.410.[28] The increase in the standard deviation across Monte Carlo replications indicates that there is a sizable loss in efficiency. This is because one is unsure of the true functional form. However, the sizes of the tests in the over-parameterized model do appear to be approximately correct. Additionally, in 70% of the replications one would reject the null hypothesis that there is no impact of the community-level variable on the outcome. The highly flexible model removes the

[28] A fully interacted quadratic model has average estimate of the mean derivative of 1.024 with standard deviation 0.266.

bias from the simple, under parameterized model without becoming extremely imprecise even in these much smaller sized samples.

Standard Error Estimators in Multilevel Models with More Than Two Levels

The analysis we reported on above indicated that one could obtain correct inferences from ordinary least squares model estimates that ignored the multilevel error structure provided one adjusted the standard errors to ex post account for the within community error correlation. The approach we used to adjust the standard errors of the estimates allowed for arbitrary forms of heteroscedasticity and error correlation within communities, but we only examined the performance of the standard error estimators when there were two levels in the analysis. It could be the case, for example, that individuals live in families which reside in communities, and there may be determinants of the individuals' behaviors that depend on unobserved family characteristics as well as unobserved individual and community characteristics In this section we consider the performance of standard error estimators when the error term has up to three levels.

Extending the descriptive notation used above, the three levels in this model are the individual level, the family level, and the community level. The DGPs we consider have the individual-level outcome being influenced by one explanatory variable from each level. Let $X_C(c)$ be the community-level explanatory variable, $X_F(f,c)$ be the family level variable, and $X_I(i,f,c)$ be the individual-level variable. We allow these explanatory variables to be correlated within communities and families, and we set $Cor[X_C(c), X_F(f,c)]=Cor[X_F(f,c), X_I(i,f,c)]=0.5$ and $Cor[X_C(c), X_I(i,f,c)]=0.667$. We permit there to be unobservable determinants of the individual-level outcomes associated with each of these three levels. The linear regression model we examine takes the following form:

$$Y(i, f, c) = \boldsymbol{b}_C \cdot X_C(c) + \boldsymbol{b}_F \cdot X_F(f, c) + \boldsymbol{b}_I \cdot X_I(i, f, c) + \mathrm{E}_T(i, f, c)\ ,$$

where

$$\mathrm{E}_T(i, f, c) = \boldsymbol{r}_C \cdot \mathrm{E}_c(c) + \boldsymbol{r}_F \cdot \mathrm{E}_F(f, c) + \boldsymbol{r}_I \cdot \mathrm{E}_I(i, f, c)$$

${}_C$(c) gives rise to the within level 3 error correlation (community), ${}_F$(f,c) gives rise to level-two error correlation (family), and ${}_I$(i,f,c) is the level-one error term (individual). The ${}_C$(c) are independent across different communities (level-three observations), the ${}_F$(f,c) are independent across families (level-two observations), and the ${}_I$(i,f,c) are independent across all individuals. These three error components are distributed as independent N(0,1) random variables. We set ρ_C, ρ_F, and ρ_I to achieve different correlation patterns for the error terms and R^2 values. In the Monte Carlo simulations we set the three regression coefficients equal to 1.0 (i.e., $\beta_C = \beta_F = \beta_I = 1$). We specify four level-one units (individuals) within each of 25 level-two units (families) for each of 200 level-three units (communities), for a total of 20,000 individual-level observations.

Our primary concern here is how one can carry out unbiased tests in these three-level models. Figures 7, 8, and 9 contain pertinent information about the size performance of various estimators of standard errors for different configurations of the multilevel error correlations. The top row of graphs in each of these three figures display information for hypothesis tests about the impact of the community-level (level 3) variable. The second row of graphs presents similar information for the impact of a family-level (level 2) variable that is correlated with the community-level variable. The third row of graphs presents the same information but for the impact of an individual-level variable that is correlated with both the family- and the community-level explanatory variable.

The left-hand side graphs examine various ways to estimate standard errors for the OLS point estimators. We consider three standard error estimators for these OLS estimates. The first is the naive standard errors as reported by standard OLS procedures assuming completely uncorrelated disturbances (labeled olstest). The second is an Eicker- Huber-White standard error estimator assuming that only observations within the second level are correlated (labeled olshfam). These standard error estimators would be appropriate, for example, if there could be non-zero error correlation among individuals within the same family ($\rho_F \neq 0$) but no correlation of disturbances across families living within the same community ($\rho_C=0$),. The third standard error estimator is similar to the second, except that it allows for possible error correlation at the third level among level-two units (e.g., error correlation among families and individuals living within the same community, labeled olshcom).

The right-hand side graphs are based on maximum likelihood point and standard error estimators that naively assume a two-level error hierarchy.[29] The first assumes that all level-one observations are equally correlated within the level-three units (labeled mlecomm). This would be the case, for example, if community-level unobserved factors could influence an individual's outcomes($\rho_C \neq 0$), but there are no unobserved family-level factors influencing the individual-level outcome ($\rho_F=0$). The second set of maximum likelihood point and standard error estimators

[29]What we attempt to evaluate here are simple-to-use estimators that are available in many multi-purpose statistical packages. Consequently we do not examine correctly specified maximum likelihood estimators that recognize the possible three level error structure. Such models should provide accurate estimates and unbiased hypothesis tests.

assumes that there is only error correlation among level-one units within the same level-two unit (e.g., only disturbances for individuals within the same family are correlated, i.e., $\rho_C = 0$ and $\rho_F \neq 0$, labeled mlefam).

The graphs display the empirical Type I error (size) for null hypotheses of the form $H_0: \beta = \beta_0 \; vs \; H_1: \beta \neq \beta_0$, where β_0 is the true value of the parameter in the DGP (i.e., 1.00 for all parameters examined), as a function of the intraclass correlation coefficient among individuals at level one within each level-two unit.[30] Each of these tests take place at a five percent level, and we carry out each test for each of the 1000 Monte Carlo replications. As in the analysis of standard error estimators in the simpler models, a point on the graph represents the fraction of times the true null hypothesis is rejected using that particular point and standard error estimator at the specified level of the intraclass correlation. An accurate standard error estimator for a particular point estimator would exhibit a straight, horizontal line at 0.05 for all values of the intraclass error correlation. Note that the vertical scales vary across graphs within these figures.[31]

Figure 7 considers the case where there is only error correlation among level-one units within the same level-two unit (e.g., only error correlation among individuals within the same family). In particular, $\rho_C = 0$, while $\rho_F \neq 0$.[32] Looking first at tests on the impact of the

---

[30]If individuals are members of families located within communities, then the intraclass correlation we consider is the correlation of the disturbances among individuals within the same family.

[31]Figures 7, 8, and 9 only examine tests at size 0.05. We obtained quite similar results for size 0.10. We set the overall error variance for the Monte Carlo experiments summarized in these figures to achieve an R^2 of 0.10. The size comparisons do not depend on the R^2 value.

[32]For Figure 7, we define the error term for each level-one observation as proportional to $\mathrm{E}^*_T(i,c,f) = \rho \mathrm{E}_F(c,f) + \sqrt{1-\rho^2}\,\mathrm{E}_I(i,c,f)$ where ρ is the level of the intraclass correlation. The two

community-level variable (level-three covariate), we see that the naive standard error estimator for the OLS point estimator performs quite poorly (olstest) . The empirical size exceeds twice the specified size even for some intraclass correlations below 0.50, with the empirical size rising to about 0.30 at the highest levels of intraclass correlation. Both of the robust, Eicker-Huber-White standard error estimators yield tests of the correct size. It is important to recognize that for these robust standard error estimators to perform correctly, one only needs to specify the highest level at which there could be error correlations.[33] Hence, the estimator allowing there to be correlations among all individuals within the same community (olshcom) provides unbiased hypothesis tests, even though there is no community-level (level 3) error correlation. Its assumption of clustering up to as high as the community level (level 3) incorporates as a special case clustering only within families (level 2). For this standard error estimator, there need not be the same form of error correlation for all observations within the level specified as being the highest level within which observations are not independent.

The maximum likelihood estimator does not generalize this way. The right-hand graph in the top row of Figure 7 indicates that the maximum likelihood estimator that models the within level two (family) correlation does provide unbiased tests; this estimation procedure coincides

---

error components are independently distributed N(0,1) random variables.

[33]There could be a cost of specifying the “clustering” level higher than is necessary. It is important for there to be enough independent higher level observations for this estimator to work well. In fact, the estimator will not provide a positive definite covariance matrix unless there are at least as many independent higher level units as parameters being estimated. Typically, one would like to have many more than this number of observations in order to obtain accurate estimators of the standard errors of the parameter estimates. If there is no community level error correlation but one specifies that there could be error correlation within communities, this will yield valid standard error estimators as long as the number of communities is large. But, if there are only a few communities the estimators might not work well.

with this specification of the DGP (only level-two correlation). The maximum likelihood estimator assuming only the higher level, community-level error correlations, however, does appear increasingly biased as the level of the intraclass correlation rises; this estimation method does not contain Figure 7's DGP as a special case. But for ρ less than 0.50, this bias appears quite small. Even at the highest levels of ρ the incorrectly specified maximum likelihood estimator provides tests that reject at most about eight percent of the time when the requested size is five percent.[34]

Turning to the estimates of the impact of the family variable (level-two covariate) in the second row of graphs in Figure 7, we find qualitatively the same results. Tests based on the naive, simple OLS standard error estimator are quite biased. The Eicker-Huber-White standard error estimators provide tests with the correct size regardless of whether one allows the highest level of error correlation to be at level two (e.g., the family) or level three (e.g., the community). The two-level maximum likelihood estimator specifying that all of the error correlation takes place at level two, which coincides with the DGP used for Figure 7, provides unbiased tests. The two-level maximum likelihood procedure specifying that the error correlation takes place at level three and not at level two provides much more biased tests than those of the impact of the community-level variable. For tests about the impact of the individual-level variable (level-one covariate), all of the standard error estimators provide unbiased tests.

Figure 8 proves the same information as Figure 7, but the DGP used for Figure 8 has all of the error correlation taking place at the community (third) level. Here $\rho_C > 0$, while $\rho_F = 0$. After

[34]We obtained approximately the same probabilities of false rejections for all approaches and for all data generating processes when we examined R^2 values of 0.90 instead of the 0.10 examined in these figures.

controlling for the level-three correlation (e.g., community), there is no additional correlation among level-one units (e.g., individuals) within the same level-two unit (e.g., families).[35] The performance of the testing procedures in this instance are somewhat different that those discussed for Figure 7. Consider first tests about the effect of the impact of the community-level variable (level-three covariate) in the top row of Figure 8. The naive OLS standard error estimator continues to provide biased tests. The Eicker-Huber-White standard error estimator with only family level (level two) error correlation and the maximum likelihood procedure that allows for error correlation only at the family level (level two) now provide quite biased tests; this is because these procedures do not recognize the level-three error correlation. The two-level maximum likelihood model that allow there to be error correlation at the community level (level three), not surprisingly, provides unbiased tests for the impact of the community-level covariate because it is correctly specified. Tests using the OLS point estimates along with the Eicker-Huber-White standard error estimators allowing for up to level-three error correlation also provide unbiased tests for the impact of the community-level covariate. Looking at the tests about the impacts of the family- and the individual-level variables (levels two and one covariates), only tests based on the naive OLS standard error estimator for the family (level two) variable are biased. All other tests appear to have the correct size.

[35]For Figure 8, we define the error term for each level-one observation as proportional to $\mathrm{E}^*_T(i,c,f) = \rho\,\mathrm{E}_C(c) + \sqrt{1-\rho^2}\,\mathrm{E}_I(i,c,f)$ where ρ is the level of the intraclass correlation. The two error components are independently distributed N(0,1) random variables.

Figure 9 presents the perhaps more realistic case when there are error correlations at level two (within the family) and at level three (within the community).[36] Looking first at the community variable (level-three covariate), the OLS standard error estimator with possible within community error (level three) correlations and the maximum likelihood approach that allow for error correlation at the community level (level three) provide unbiased tests. It is somewhat surprising that this maximum likelihood estimator performs correctly here. It performed somewhat poorly when there was only family level (level two) error correlation as in Figure 7, while here there is family error correlation as well as community error correlation. For the community-level variable effect, any approach that does not recognize that there can be correlated errors at the community level provides biased hypothesis tests. For the family-level variable (level-two covariate, second row of graphs in Figure 9), the maximum likelihood approach that allows for only community-level (level three) error correlation now performs poorly. The naive OLS approach continues to provide biased tests. The maximum likelihood approach that allows for only level-two (family) error correlation, and the Eicker-Huber-White standard error estimator assuming independence above level two, provide unbiased tests for this family-level variable, even though they both performed poorly for tests about the community-level variable. The Eicker-Huber-White standard error estimator that allows for arbitrary community-level (level three) error correlation continues to provide unbiased tests. For the individual-level variable all approaches

[36]For Figure 9, we define the error term for each level one observation as proportional to $E^*_T(i,c,f) = r\sqrt{0.5}[E_C(c) + E_F(c,f)] + \sqrt{1 - r^2}E_I(i,c,f)$ where ρ is the level of the intraclass correlation. The three error components are independently distributed N(0,1) random variables.

provided unbiased tests, even the naive OLS approach that assumes all observations are uncorrelated.

Overall, the main conclusion about the performance of standard error estimators when there are three-level models is that it is most important to control for the error correlation at the highest possible level at which it might exist. For the OLS estimates with Eicker-Huber-White standard errors, it does not matter whether one “over-controls” and allows for possible error correlations at a higher level than is actually the case. For these Eicker-Huber-White standard error estimators, as long as the highest level of actual error correlation is nested within the level specified in the estimation, hypothesis tests will be unbiased. In fact, the only cost of specifying possible correlations at too high a level for the Eicker-Huber-White estimators is that the standard error estimators might become imprecisely estimated if there are too few observations at the highest level specified. For the two-level maximum likelihood estimator, it is important to specify exactly the level at which the error correlation takes place.

There are some instances where tests based on the incorrectly specified maximum likelihood procedures perform well. For the most part it appears that hypothesis tests about level-three (community) variables will be at most only slightly biased if one assumes that the error correlations take place only at level three (community level). It appears that tests about the impacts of level-one (individual) variables based upon the maximum likelihood procedures assuming only error correlations at level two (family) are unbiased. If one is going to use two-level maximum likelihood estimators in the presence of three level error components, these results suggest it would be best to assume that all error correlation takes place at the highest level (e.g., community level).

It is important to note that all of the Monte Carlo experiments we report on here allow there to be correlations among the explanatory variables only through the observed level-two (family) and level-three (community) explanatory variables. If there are other reasons why there could be correlations of regressors across level-one and level-two units, then neither of the simple two-level maximum likelihood procedures nor the Eicker-Huber-White standard errors allowing only correlations within level one would provide unbiased tests for the impacts of the individual-level (level one) covariates. In this instance only the Eicker-Huber-White standard errors allowing for possible correlations within level three provides unbiased tests.

How Do You Know if You Need to Control for the Multilevel Error Structure?

The simple answer to this question is that if you at all suspect that there could be unobserved determinants from a higher level then you should use standard error estimators that recognize the possibility of error correlation within lower level units. Typically, if one fails to recognize the possibility of correlated disturbances then most hypothesis tests will reject too frequently true null hypotheses. The Eicker-Huber-White standard error estimators as implemented in Stata provide unbiased tests even if one "over-specifies" the highest level within which there could be error correlation. In a sense, it is costless to use these adjusted standard errors. If one adjusts the standard errors unnecessarily, the adjustments will tend to be quite small and inconsequential. As an added benefit, these Eicker-Huber-White standard error estimators also control for arbitrary forms of heteroscedasticity. Heteroscedasticity is a key consequence of having random coefficient models, so this type of standard error adjustment could be quite useful for a variety of reasons. Therefore, the standard errors of simple OLS point estimates should be adjusted.

The harder question to answer is whether one should implement maximum likelihood procedures to control for the multilevel error structure. Before answering this question, it is important to recognize the potential gains from using a maximum likelihood multilevel model.[37] The gain in efficiency from using models that incorporate the multilevel structure varies according to the type of covariate. For covariates measured at the highest level, typically the efficiency gains

[37]Throughout this discussion we assume either that all parameters are fixed (i.e., effects of covariates do not vary across individuals) or that one is interested only in the mean parameter value if parameters are random (e.g., vary by individual or by community, or both, in the two-level models). An evaluation of the performance of multilevel, random parameter models is well beyond the scope of this essay. However, it would be quite useful for there to be detailed evaluations of these more complicated estimation procedures that are similar to those we present here for the simple multilevel model.

for maximum likelihood estimators over OLS estimators are small. If one is interested in estimating the impact of a community-level program on individual-level outcomes, then there is little to gain. For estimating the impacts of individual-level covariates in individual-level outcomes, however, the efficiency gains can be substantial.

There are two possible costs of using multilevel models. The first is that they are slightly more difficult to estimate than simple OLS models. For simple models this cost should be inconsequential, as several standard statistical packages do incorporate seemingly unrelated regression models and maximum likelihood procedures for two-level error structures. The second cost of using multilevel estimation procedures is that the standard errors reported from these models will be incorrect unless the true form of the multilevel model is specified in the estimation procedure. Unlike the Eicker-Huber-White standard error estimators, the standard error estimators from the maximum likelihood models will be incorrect unless one models correctly the form of the error structure at all levels.

It should, however, be possible to minimize the importance of this latter cost. One can use the maximum likelihood point estimators, even with a somewhat misspecified error structure, and then adjust the standard errors by using Eicker-Huber-White standard error estimators adapted for “quasi-maximum” likelihood models. Unfortunately, such procedures are not readily available in existing computer packages with multilevel error structures. If they were implemented in these computer packages, then researchers would be able to carry out valid hypothesis tests while retaining some of the efficiency gains by exploiting the multilevel error structure.

The answer to the question of whether one should use multilevel models is complex. In many interesting situations, such as the estimation of community level characteristics in the presence of community-level unobservables, there will be little efficiency gains. In this instance there is no compelling reason to undertake more difficult estimation problems. In other situations, such as the estimation of the effects of individual-level variables, there can be important efficiency gains. But in these instances, it is, in general, key to specify correctly the precise form of the multilevel model in the estimation. Regardless of whether one takes the multilevel structure of the errors into account, it does seem important to use robust standard error estimators, like the Eicker-Huber-White ones used here, unless one is sure that the estimation model is correctly specified.

Conclusion

In this essay we have explored several issues about the importance of using multilevel modeling approaches when analyzing data of the type frequently used to evaluate health and family planning programs in developing countries. In particular, we examined a simple model where an individual-level outcome could depend on individual-level covariates and on covariates that come from a higher, more aggregate level. The salient feature that makes these models more complex than standard regression models arises from the fact that there may be unobserved determinants of behavior at both the individual and at the more aggregate level. In this instance, several observations within the same aggregate level will be influenced by the same higher level unobserved determinants and consequently will have correlated disturbances. This violates one of the key assumptions for the Ordinary Least Squares estimators to be the Best Linear Unbiased Estimator; it also violates a key assumption for the usual OLS formulae to provide accurate and reliable standard error estimators. There are maximum likelihood estimators for these multilevel models that recognize these error correlations and exploit them to obtain theoretically better estimators. We gauge the performance of the OLS estimators against these maximum likelihood estimators that theoretically provide the most precise, unbiased estimators.

Theoretically, the presence of multilevel, correlated disturbances does not introduce a bias into the estimated parameters for any of the estimators considered here, and we demonstrate this through our first set of Monte Carlo experiments. While bias of parameter estimates is not an issue, there are two drawbacks to ignoring multilevel, hierarchical disturbances and relying on simple OLS procedures to evaluate the impacts of covariates on outcomes. The first limitation of the OLS estimators is that one might be able to define more accurate estimators of the impacts of the individual- and aggregate-level variables on the individual-level outcome. The second is that the standard errors reported by ordinary least squares procedures will not reflect accurately the true sampling variability of the estimators. This latter shortcoming can lead to biased tests of hypotheses about the parameters of interest; nearly always the standard errors reported by the OLS procedures will be too small, and researchers will reject true null hypotheses more frequently than they specify for the sizes of their tests.

Our analysis reveals that there are only very small efficiency gains to be obtained from using the most efficient estimators of the impact of the higher level (e.g., community level) variable on the lower level (e.g., individual level) outcome instead of Ordinary Least Squares estimators. Theoretically, all efficiency gains for this coefficient estimator from using the maximum likelihood estimators instead of the OLS estimator are the result of the higher level explanatory variable being correlated with the lower level explanatory variable. This means that any potential efficiency gains are tempered by the inclusion of highly correlated variables which cause the precision of the estimator to fall. Only if there is both a high correlation of the regressors from the different levels (e.g., greater than 0.90) and a high correlation of disturbances within the higher level units can the efficiency improvements for the impact of higher level explanatory variable be at all substantial. There can be somewhat larger efficiency gains when there are only a few observations per higher level unit (i.e., few individuals per community), but these diminish rapidly after having only five or ten observations per higher level unit. But even with two or three observations per community, the gains are substantively small unless there are quite high correlations between the regressors at the two levels. There can, however, be substantial efficiency gains for the estimators of the impacts of the lower level (individual level) variables on the lower level (individual) outcome, but such impacts are typically only of second order interest when evaluating the impacts of programs on individual outcomes.

Even though the OLS point estimators appear to perform quite well relative to the maximum likelihood estimators in most applied situations, the standard error estimators provided by standard Ordinary Least Squares formulae are incorrect in the presence of multilevel error correlations. For two-level models, we find that the robust asymptotic approximations to the standard errors of the OLS model due to Eicker, Huber, and White provide approximately unbiased tests for all parameter estimators when one uses formulae that allow error correlations at the higher level. The maximum likelihood standard error estimators perform flawlessly for these two-level models.

When we examine three-level models, the Eicker-Huber-White standard error estimators allowing for error correlations within the highest level continue to perform quite well, while the maximum likelihood estimators that assume only two levels often perform poorly. This failure of the maximum likelihood estimators is due to the fact that they are incorrectly specified for the three-level models we examine. It is important to note that one will usually obtain biased tests with these "maximum likelihood" estimators even though they control for correlations at the highest level. This could be an important factor to consider when using maximum likelihood estimators if there could be a missing "middle" level in a researcher's empirical model; the Eicker-Huber-White standard error estimators do not have this limitation.

If one is primarily interested in estimating the impacts of a community-level variable on individual-level outcomes, as is frequently the case in the evaluation of health and family planning programs in developing countries, then the results of this paper provide some important guidelines. First, there appear to be small efficiency gains for the estimates of the impacts of community-level factors on the individual-level behavior from using maximum likelihood procedures instead of simple ordinary least squares estimation. Second, it is crucial to adjust the estimated standard errors of the ordinary least squares estimators to reflect the fact that there can be correlated error terms at higher levels; the Eicker-Huber-White standard error estimators appear to provide adequate adjustments. Third, when one estimates incorrectly specified regression functions, multilevel models can indicate that the multilevel error structure is important, even when only simple adjustments of the regression function indicate that the multilevel models are irrelevant. This point, given the first two guidelines, suggests that it might be more important for researchers to investigate more detailed regression function specifications before they attempt to use the more complex, maximum likelihood, multilevel procedures. Fourth, even if there are complex multilevel error correlations in the data, the Eicker-Huber-

White standard error adjustments always provide unbiased tests, as long as one allows for error correlation at the highest level; simple two-level maximum likelihood models do not provide unbiased tests when lower level error correlations are present. In summary, these results indicate that simple ordinary least squares models with standard errors corrected for high level error correlation appear to provide unbiased and accurate estimates of the impacts of community-level variables on individual-level outcomes.

Figure 1

The Unbiasedness of Ordinary Least Squares and Maximum Likelihood Estimators in Models with Multilevel Errors by the Level of the Intraclass Error Correlation

Figure 1 A: Community Level Variable Coefficient Estimates at Three R^2 Values

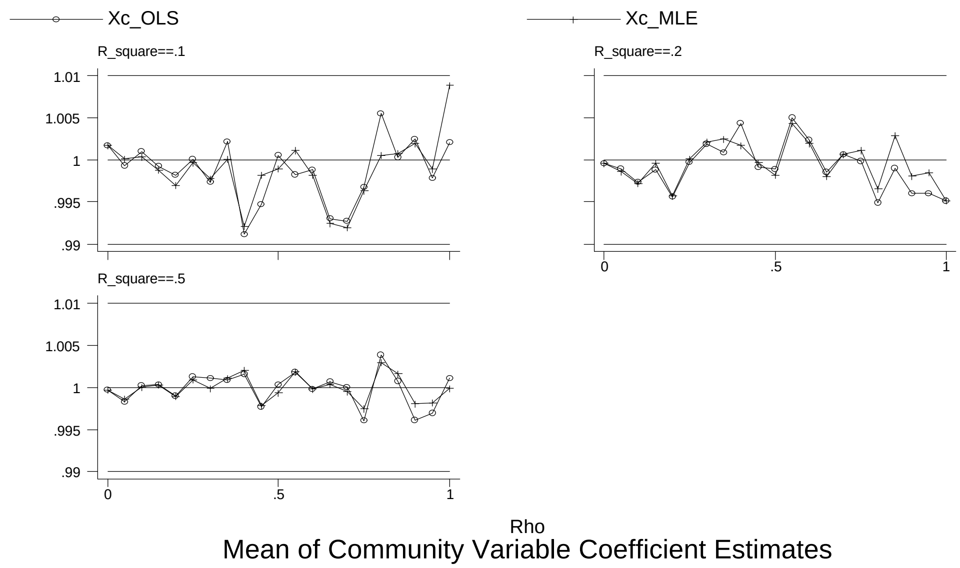


Figure 1 B: Correlated Individual Level Variable Coefficient Estimates at Three R^2 Values

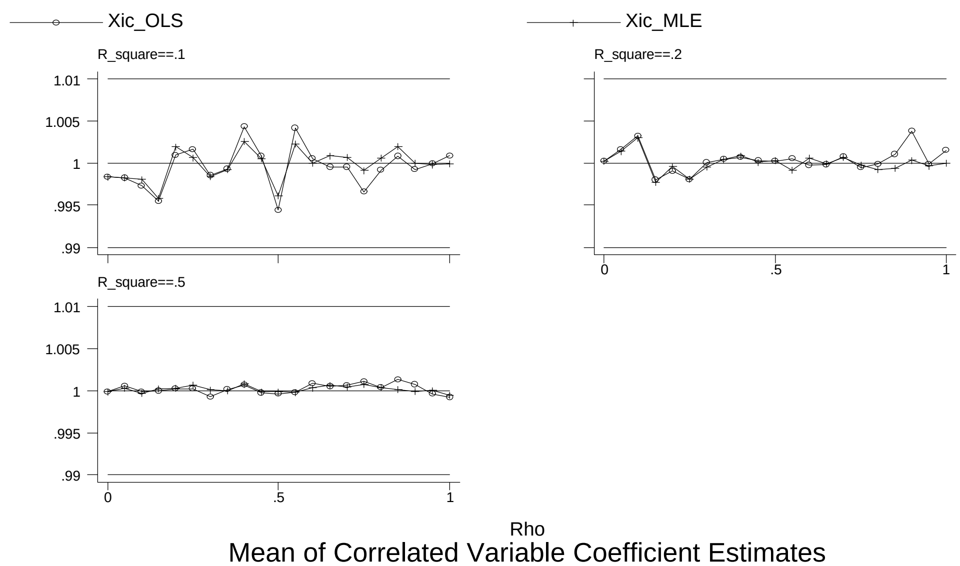


Figure 1 C: Independent Individual Level Coefficient Estimates Three R^2 Values

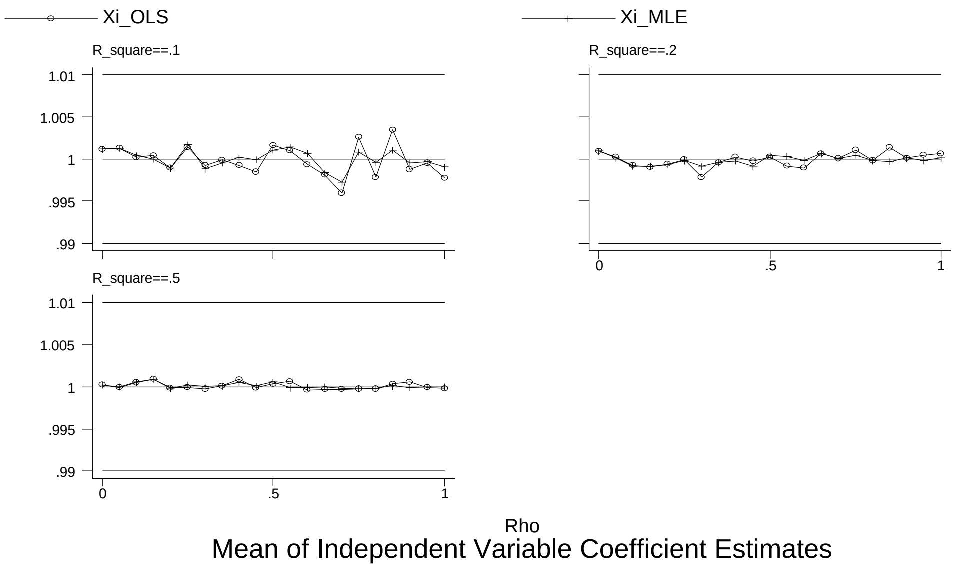
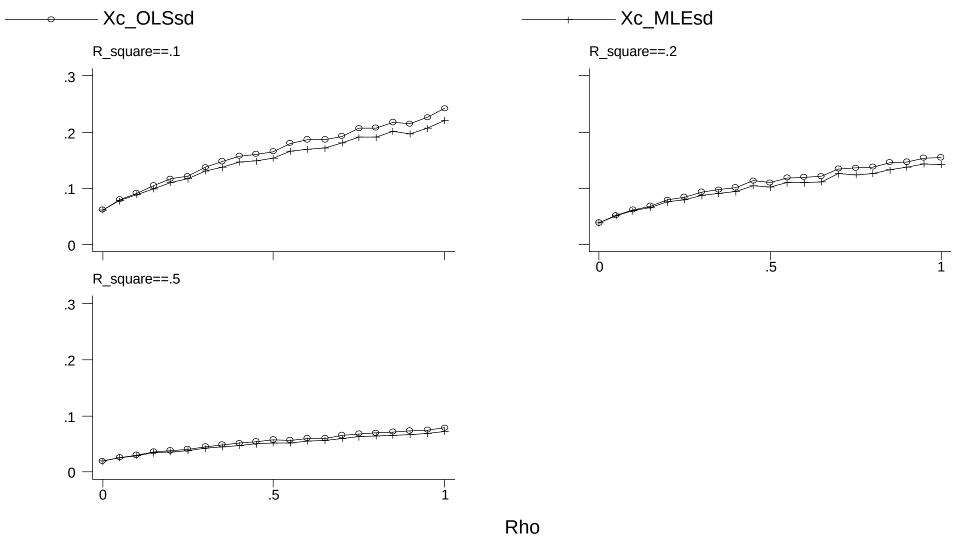


Figure 2

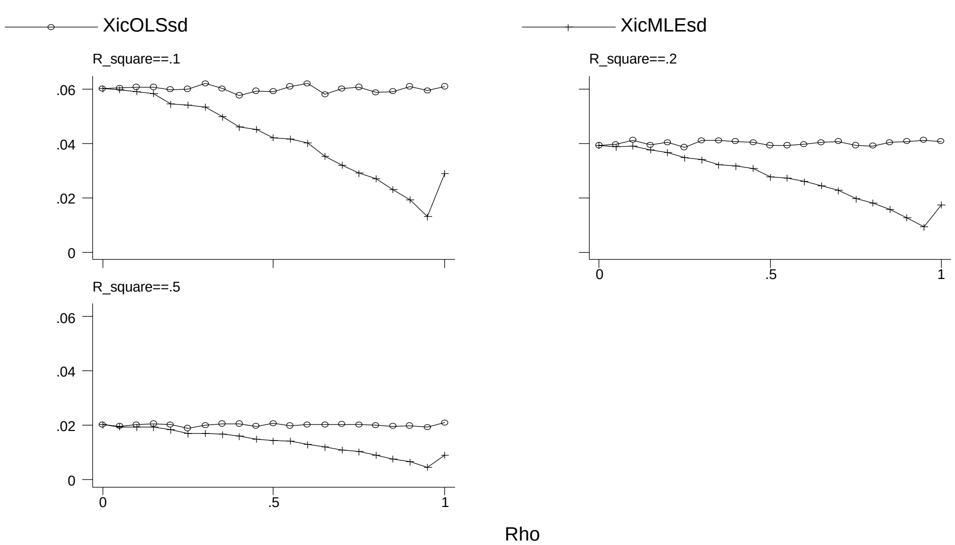
Empirical Standard Deviations of Ordinary Least Squares and Maximum Likelihood Estimates in Multilevel Models by the Level of the Intraclass Error Correlation

Figure 2A: Standard Deviations for Community Level Coefficient Estimates at Three R^2 Values



Standard Dev. of Community Var. Coefficient Estimates

Figure 2B: Standard Deviations for Correlated Individual Level Coefficient Estimates at Three R^2 Values



Std. Dev. of Correlated Individual Var. Coefficient Estimates

Figure 2C: Standard Deviations for Independent Individual Level Coefficient Estimates at Three R^2 Values

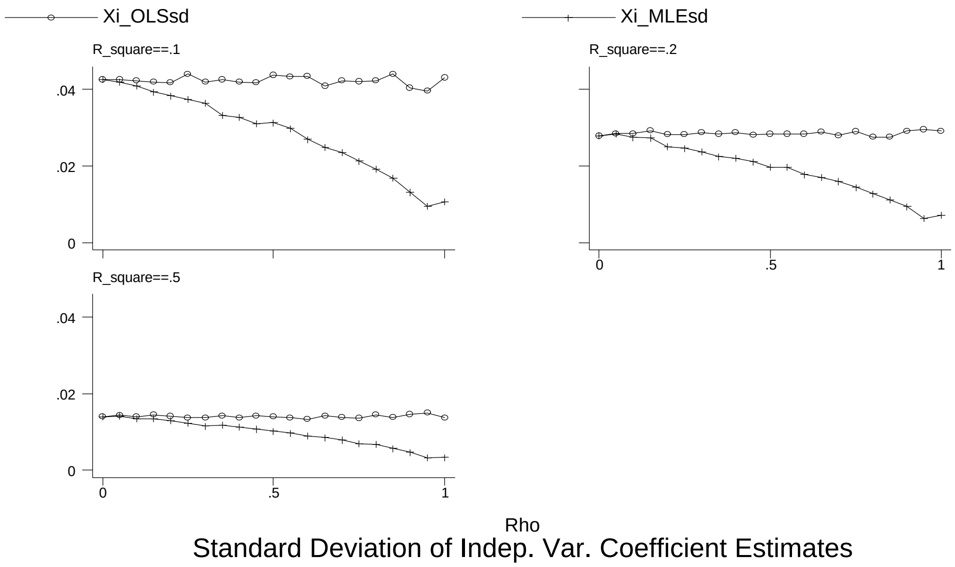


Figure 3

Standard Deviations of Ordinary Least Squares Estimators as a Fraction of the Standard Deviations of the Maximum Likelihood Estimators as a Function Number of Observations per Community

Figure 3A: Community Level, Correlated Individual Level, and Independent Individual Level Coefficient Estimators for Intraclass Correlation 0.25 and Four Regressor Correlations

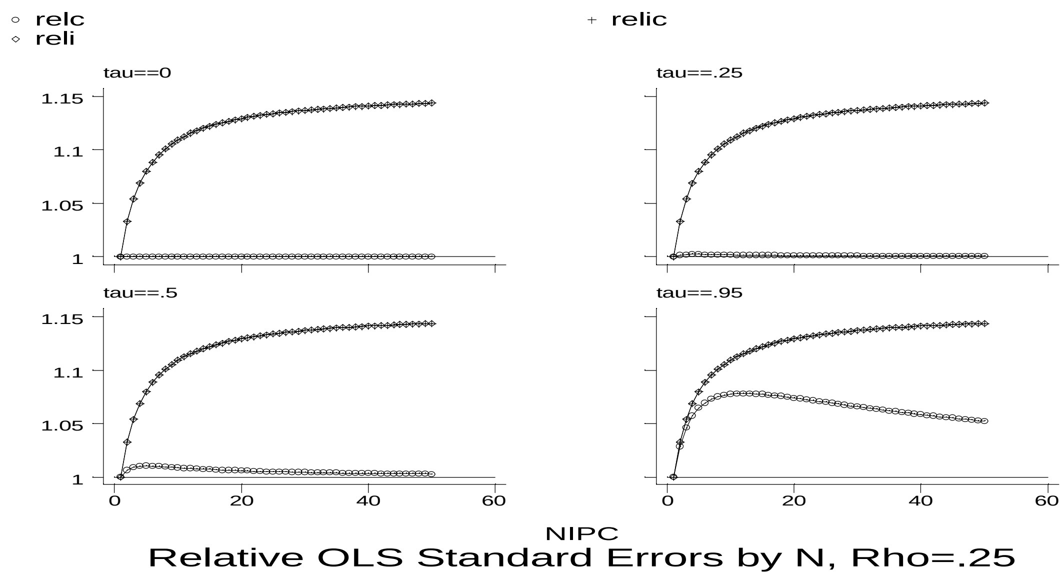


Figure3B: Community Level, Correlated Individual Level, and Independent Individual Level Coefficient Estimators for Intraclass Correlation 0.75 and Four Regressor Correlations

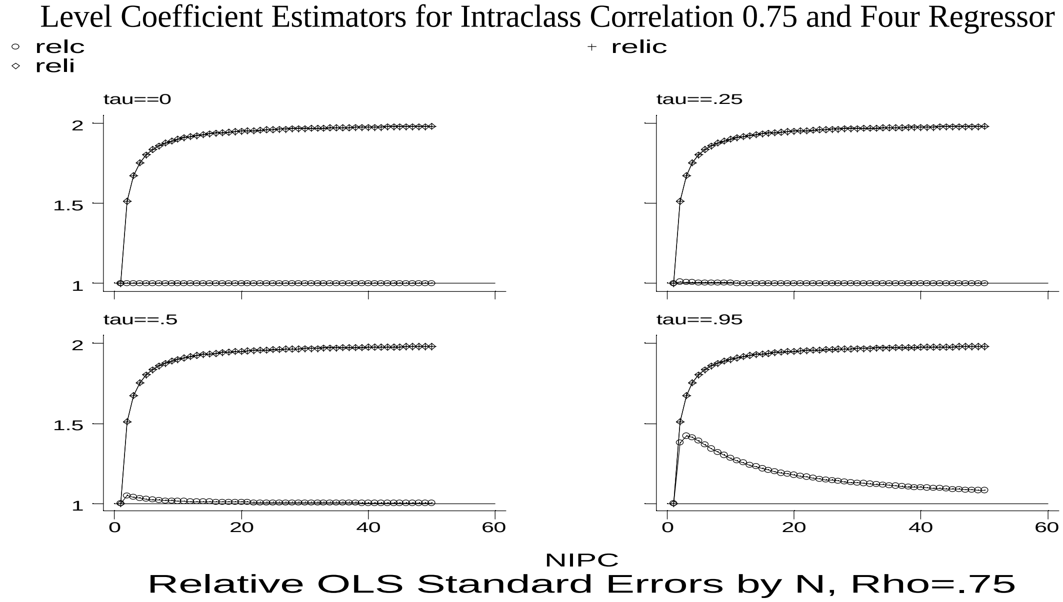


Figure 4

Standard Deviation of Ordinary Least Squares Estimator as a Fraction of the Standard Deviation of the Maximum Likelihood Estimator of the Impact of the Community Level Variable, as a Function of the Correlation of the Community and Individual Level Regressors ( )

Figure 4A: Community Level Coefficient Estimators with an Intraclass Correlation of 0.25 and Four Specifications of the Number of Observations per Community

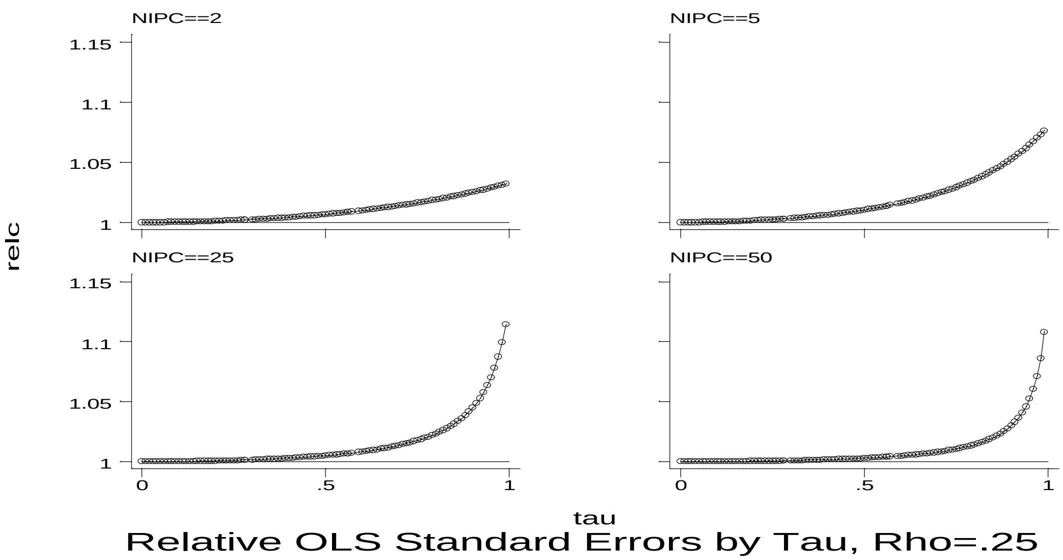


Figure 4B: Community Level Coefficient Estimators with an Intraclass Correlation of 0.75 and Four Specifications of the Number of Observations per Community

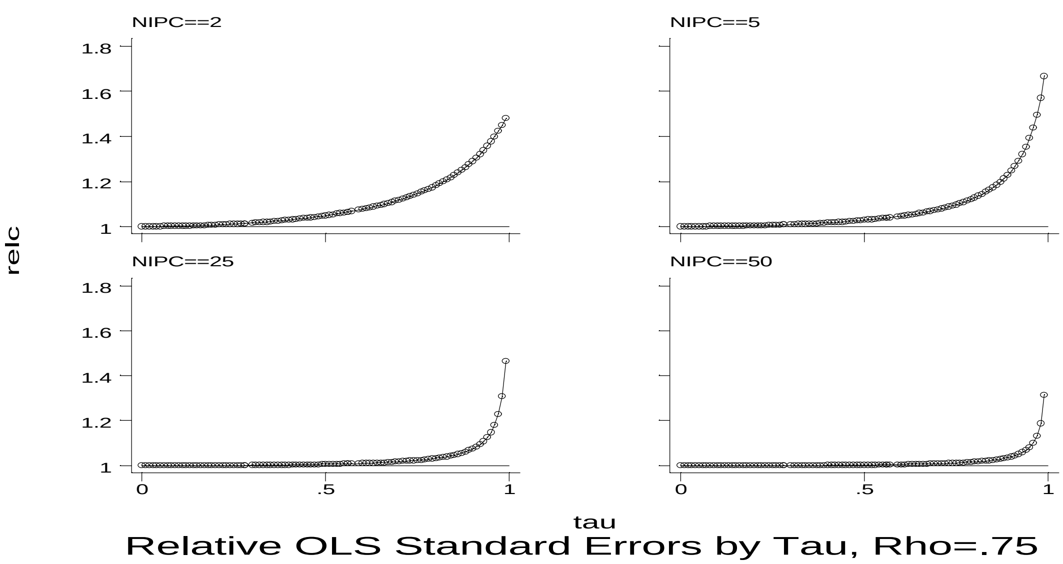
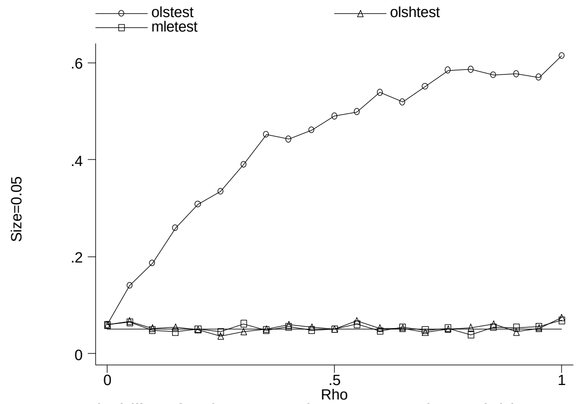


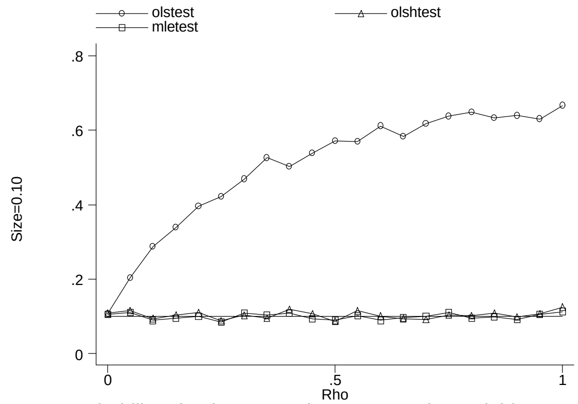
Figure 5
Performance of Standard Error Estimators: Probability of Rejecting a True Null Hypothesis

Size = 0.05 | Size = 0.10

Figure 5A: Community Level Variable

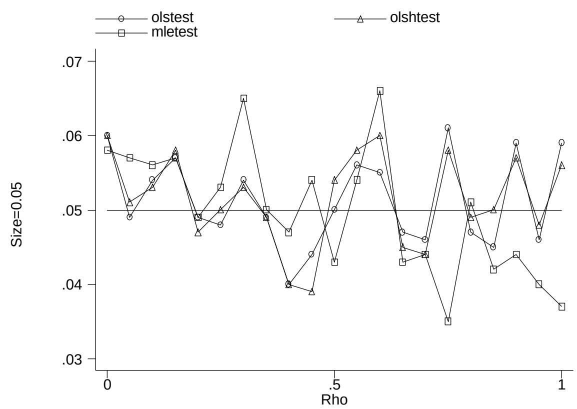


Probability of False Regection, Community Variable

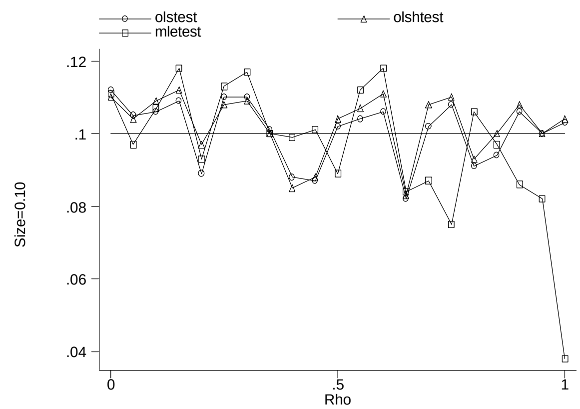


Probability of False Regection, Community Variable

Figure 5B: Correlated Individual Level Variable

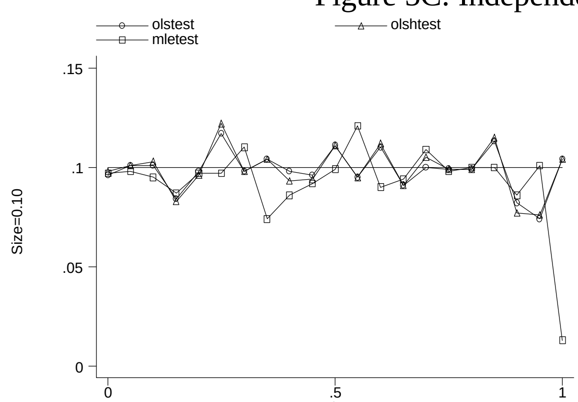


Prob of False Regection, Correlated Individual Variable

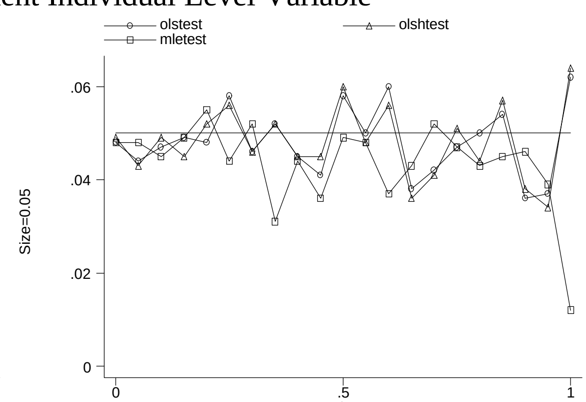


Prob of False Regection, Correlated Individual Variable

Figure 5C: Independent Individual Level Variable



Prob of False Regection, IndependentIndividual Variable



Prob of False Regection, IndependentIndividual Variable

Figure 6

Power to Reject Null Hypotheses as a Function of the Intraclass Error Correlation
Ordinary Least Squares Estimators with Eicker-Huber-White Standard Errors and
Two-Level Maximum likelihood Estimators

Figure 6A: Coefficient on the Community Level Variable

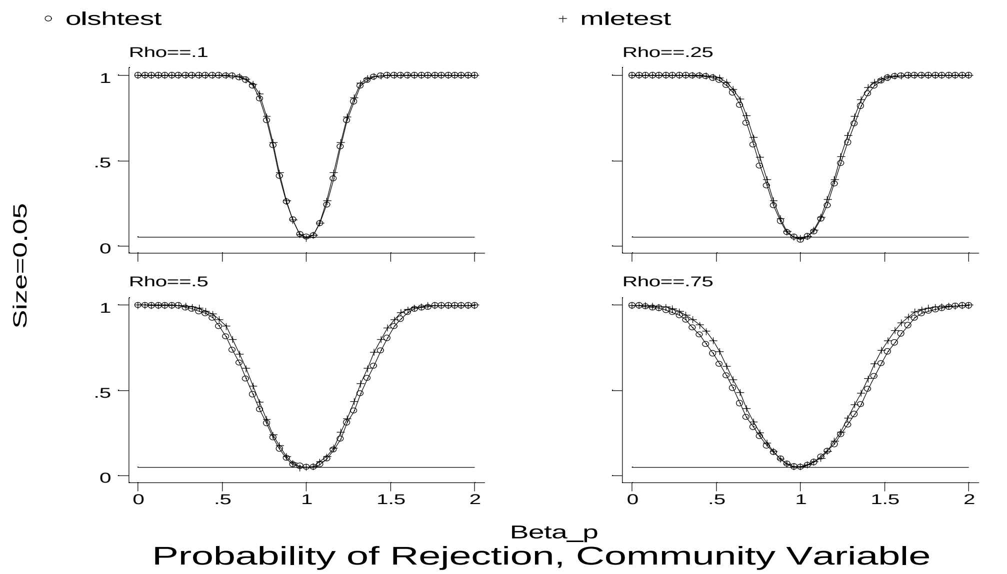


Figure 6B: Coefficient on the Correlated Individual Level Variable

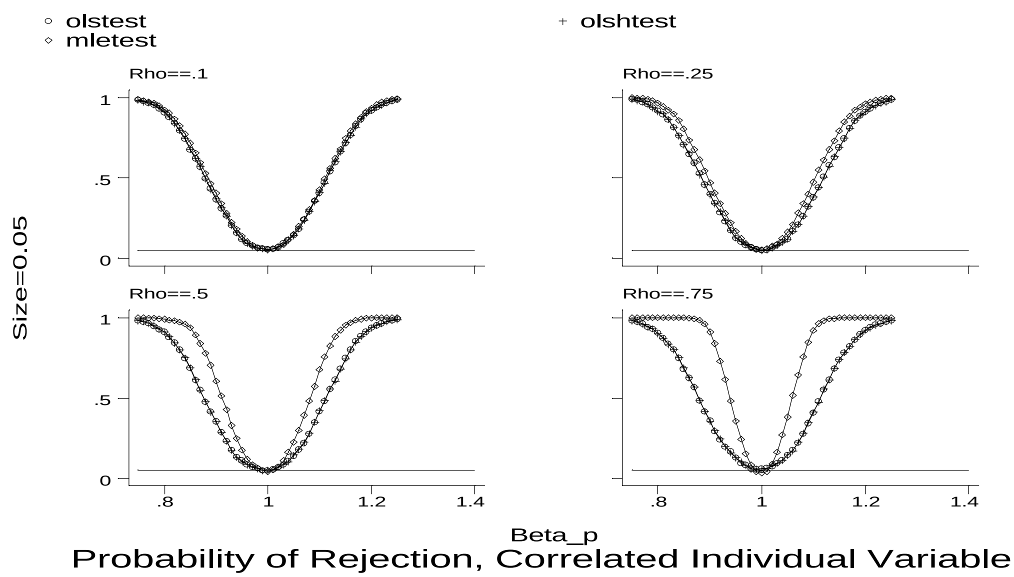


Figure 6C: Coefficient on the Independent Individual Level Variable

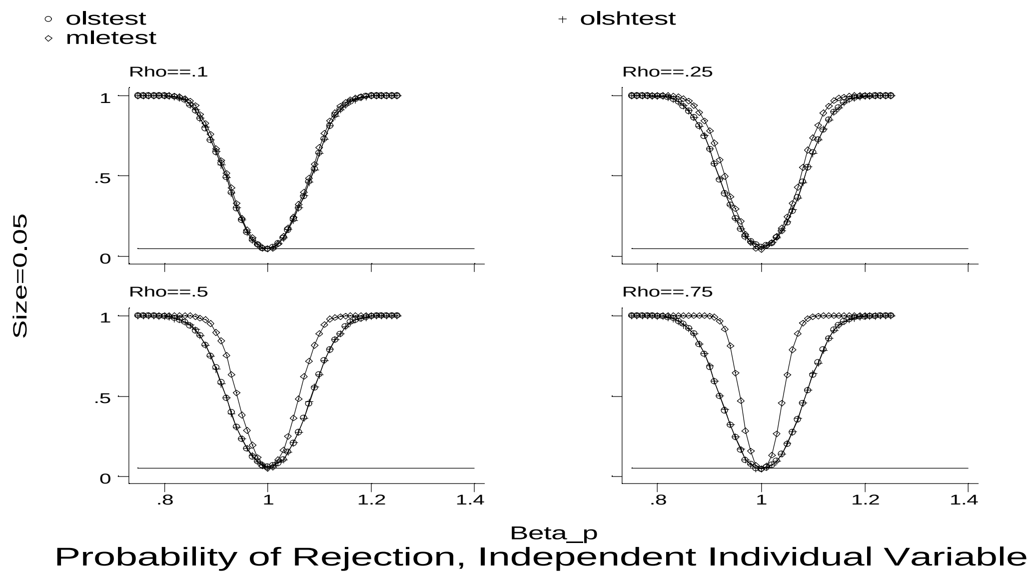
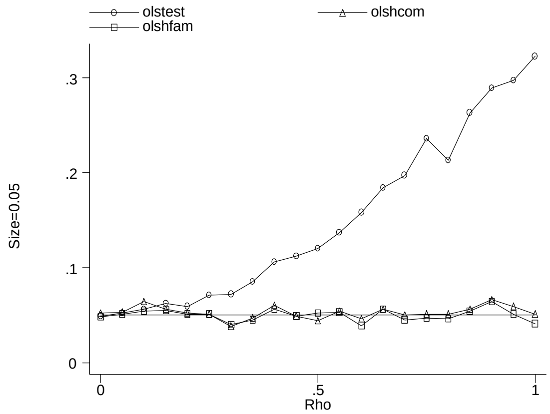


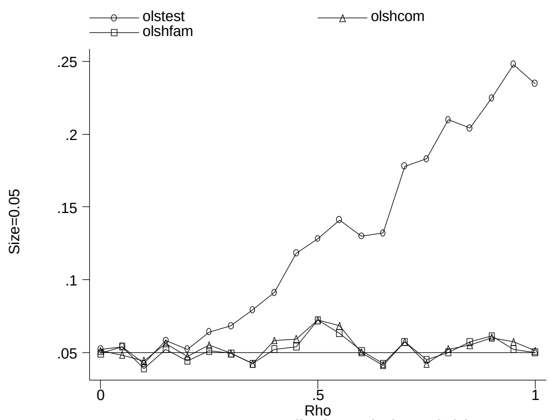
Figure 7

Performance (Size) of Standard Error Estimators for Three Level Models
Only Level Two Error Correlation

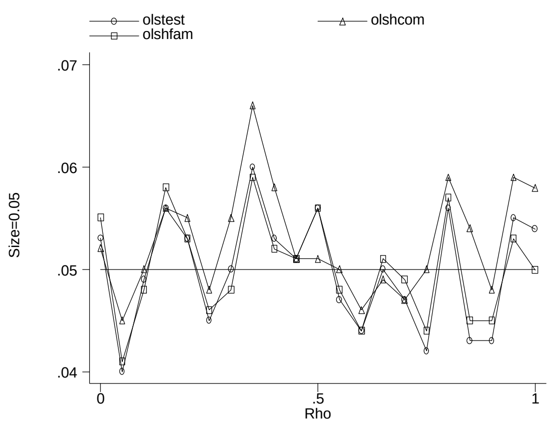
Ordinary Least Squares Estimators



Type I Error, OLS, Community (Level 3) Variable

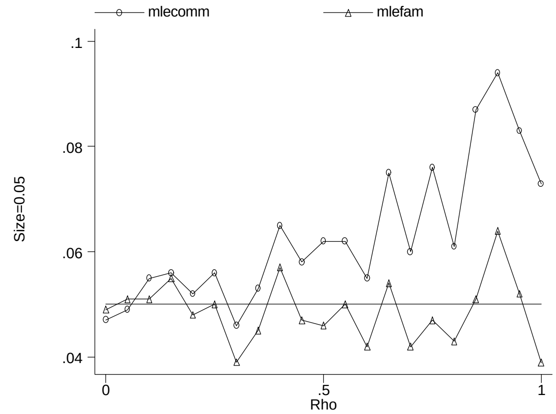


Type I Error, OLS, Family (Level 2) Variable

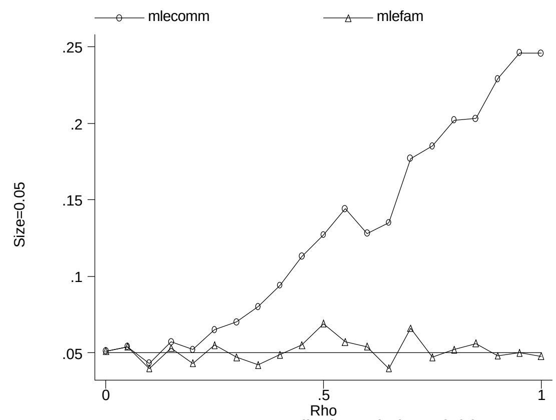


Type I Error, OLS, Individual (Level 1) Variable

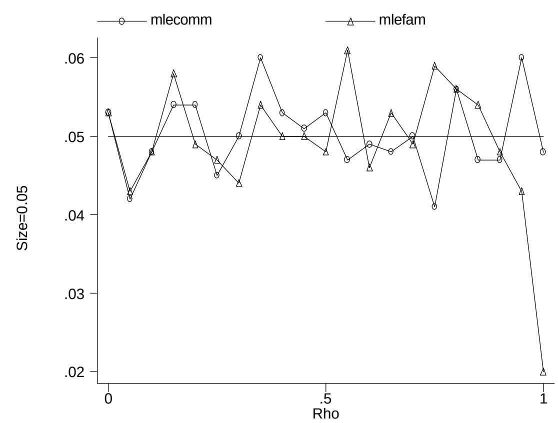
Maximum Likelihood Estimators



Type I Error, MLE, Community (Level 3) Variable



Type I Error, MLE, Family (Level 2) Variable



Type I Error, MLE, Individual (Level 1) Variable

Figure 8

Performance (Size) of Standard Error Estimators for Three Level Models
Only Level Three Error Correlation

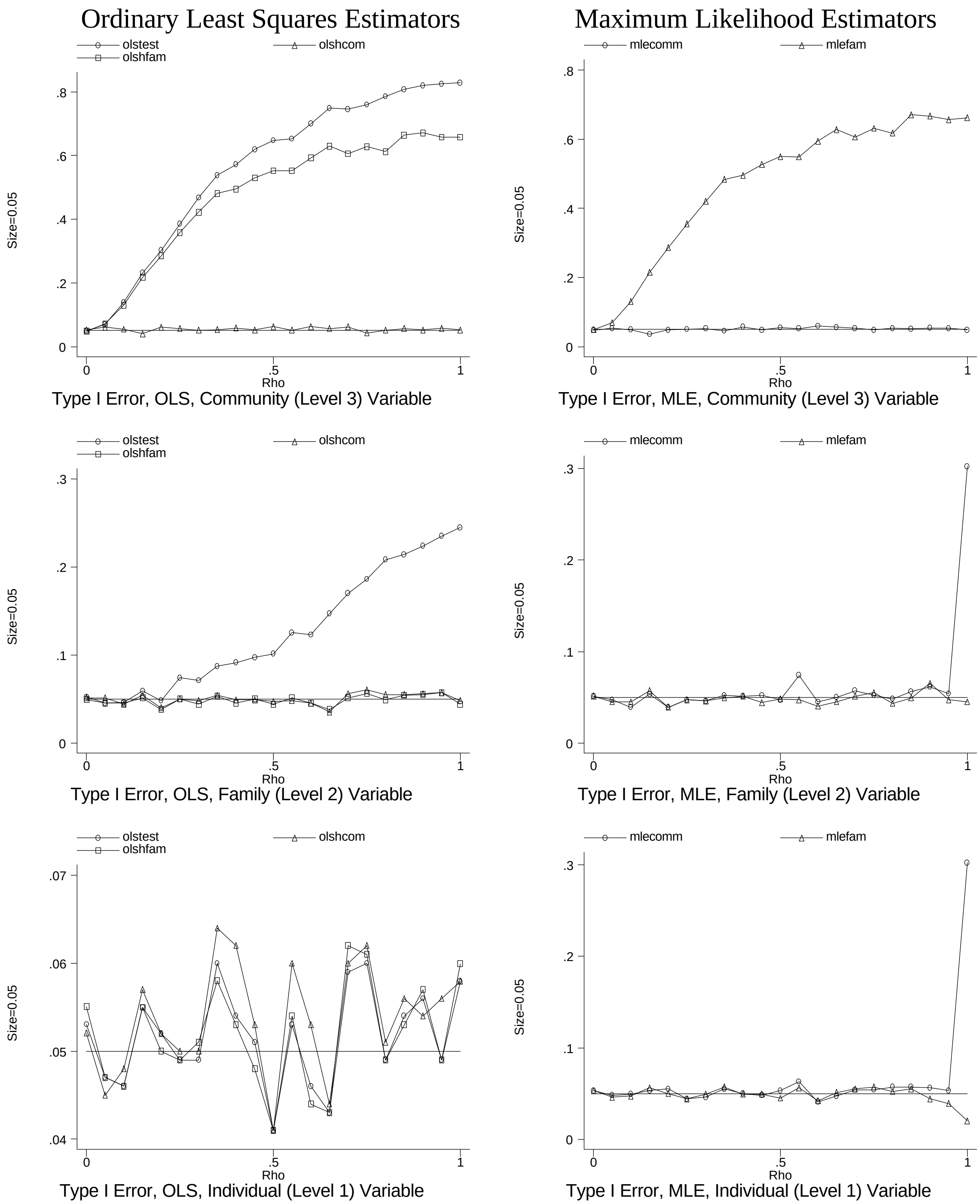
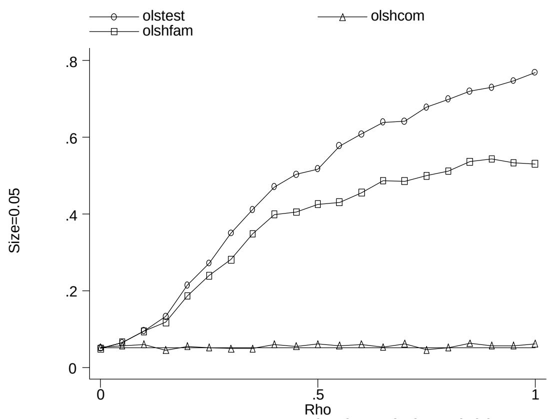


Figure 9

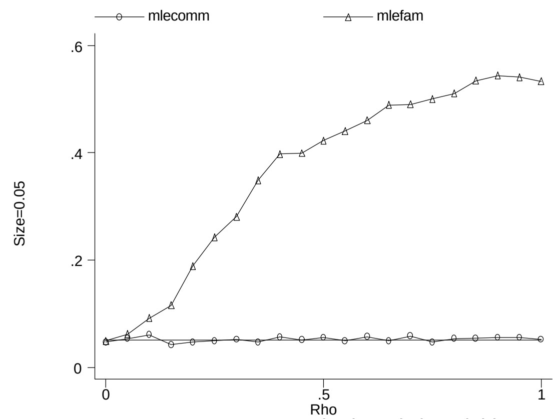
Performance (Size) of Standard Error Estimators for Three Level Models
Level Two and Level Three Error Correlations

Ordinary Least Squares Estimators

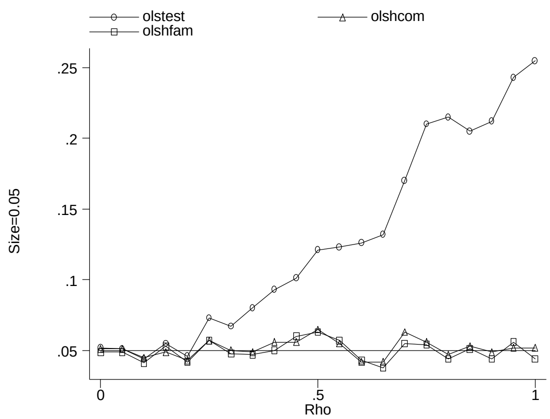
Maximum Likelihood Estimators



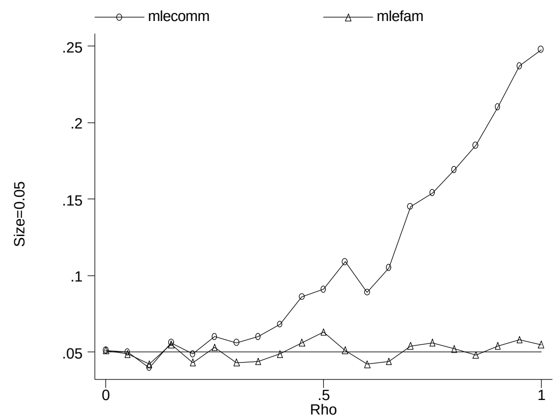
Type I Error, OLS, Community (Level 3) Variable



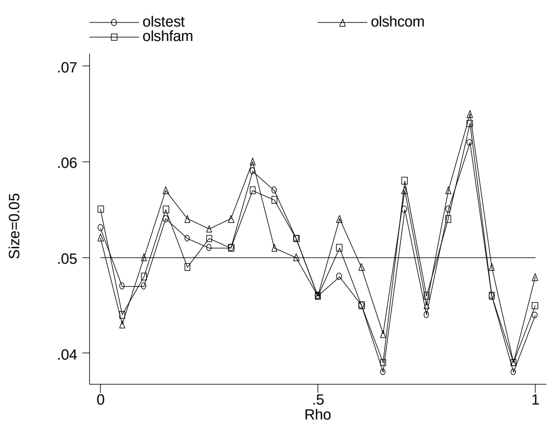
Type I Error, MLE, Community (Level 3) Variable



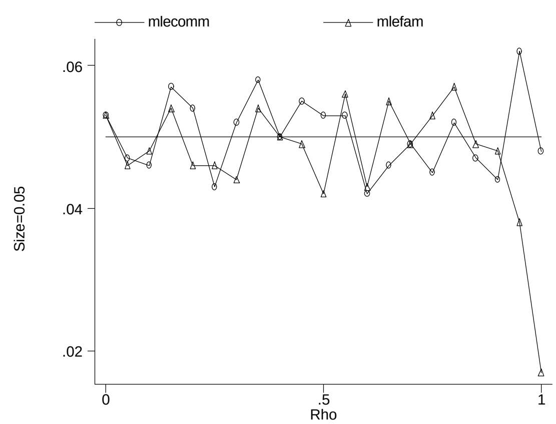
Type I Error, OLS, Family (Level 2) Variable



Type I Error, MLE, Family (Level 2) Variable



Type I Error, OLS, Individual (Level 1) Variable



Type I Error, MLE, Individual (Level 1) Variable

Appendix

Analytic Expressions for Asymptotic Covariance Matrices

Suppose there is a two level error structure in a simple regression model describing the outcome variable y(i,c).

$$Y(i,c) = b_0 + b_C X_C(c) + b_{IC} X_{IC}(i,c) + b_I X_I(i,c) \qquad \text{(A.1)}$$
$$+ \mathrm{E}_C(c) + \mathrm{E}_I(i,c)$$

We assume that there are N level 1 units (subscript i) and J level 2 units (subscript c). We assume independence across level 2 units. The disturbances $\mathrm{E}_I(i,c)$ and $\mathrm{E}_{IC}(c)$ are assumed uncorrelated and homoscedastic (equal variance not depending on the observed variables). We assume also that the explanatory variables $X_C(c)\ and\ X_{IC}(i,c)$ can be correlated, and that the correlation of the level one variables within the level two unit is due only to factors that influence the level 2 explanatory variable. In particular, if $Cov(X_I(c), X_{IC}(i,c)) = t$ then $Cov(X_{IC}(i,c), X_{IC}(i',c)) = t^2\ \forall i \neq i'\ 1$. The explanatory variable $X_I(i,c)$ is uncorrelated with both $X_C(c)\ and\ X_{IC}(i,c)$. Without loss of generality, we assume that the disturbances $\mathrm{E}_I(i,c)$ and $\mathrm{E}_{IC}(c)$ are uncorrelated and that the fraction of the total variance of

$\mathrm{E}_T(i,c) = \mathrm{E}_C(c) + \mathrm{E}_I(i,c)$ due to the level 2 disturbance, the intraclass correlation coefficient, is ρ. With only a minor loss of generality, we assume that all explanatory variables and the outcome are mean zero, so there is no intercept in the model.

Throughout the derivations in this Appendix, we assume that the variance of the composite error, $_T$(i,c), equals 1. If the composite error variance were instead ^{2}, then one would simply multiply each of the covariance matrices in this appendix by ^{2} to obtain the covariance matrices with the non-unit variance. We also assume that the variances of each of the explanatory variables equals 1. If this is not the case, then one should adjust each element of each covariance matrix by the product of the standard deviations of the explanatory variables that relate to the particular row and the particular column of the element of covariance matrix. Since we focus on ratios of the elements of the covariance matrix in this analysis, such normalizations are inconsequential.

Let X be the matrix of explanatory variables (X has NJ rows and 3 columns) and y be a column vector with NJ elements corresponding to the explanatory variables in X. The simple ordinary least squares estimator is given by

$$\hat{b} = (X'X)^{-1}X'y = b + (X'X)^{-1}X'e_T$$

where $_T$ is a NJ vector containing the composite disturbances $_T$(i,j). The covariance matrix for the OLS estimators is

$$Var(\hat{b} - b) = E[(\hat{b} - b)(\hat{b} - b)'] = E[(X'X)^{-1}X'e_Te_T'X(X'X)^{-1}]$$

Provided that the number of level one units is finite, the asymptotic distribution of the OLS estimator is

$$\sqrt{J}(\hat{b}_{OLS} - b) \xrightarrow{d} N(0, \Omega_{OLS})$$

$$where$$

$$\Omega_{OLS} = p\lim\left[\left(\frac{X'X}{J}\right)^{-1}\right] \cdot p\lim\left[\left(\frac{X'e_T e_T'X}{J}\right)\right] \cdot p\lim\left[\left(\frac{X'X}{J}\right)^{-1}\right]$$

By the properties of probability limits, and given finite fourth moments,

$$p\lim\left[\left(\frac{X'X}{J}\right)^{-1}\right]=\left[E\left(\frac{X'X}{J}\right)\right]^{-1}=$$

$$\begin{pmatrix} N & tN & 0 \\ tN & N & 0 \\ 0 & 0 & N \end{pmatrix}^{-1}=\begin{pmatrix} \frac{1}{N(1-t^2)} & \frac{-t}{N(1-t^2)} & 0 \\ \frac{-t}{N(1-t^2)} & \frac{1}{N(1-t^2)} & 0 \\ 0 & 0 & \frac{1}{N} \end{pmatrix} \qquad \text{(A.2)}$$

and

$$p\lim\left[\left(\frac{X'e_T e_T' X}{J}\right)\right]=E\left[\left(\frac{X'e_T e_T' X}{J}\right)\right]=$$

$$N\cdot\begin{pmatrix} (1+r(N-1)) & t(1+r(N-1)) & 0 \\ t(1+r(N-1)) & (1+rt^2(N-1)) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The asymptotic covariance matrix of the OLS estimates is, after some algebraic manipulation, given by

$$\Omega_{OLS} = \frac{1}{N}\begin{pmatrix} \dfrac{1+r(N-1)(1-t^2)}{(1-t^2)} & \dfrac{-t}{(1-t^2)} & 0 \\ \dfrac{-t}{(1-t^2)} & \dfrac{1}{(1-t^2)} & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{(A.3)}$$

Next consider the Generalized Least Squares (GLS), best linear unbiased estimator, that takes account of the fact that there the $_{T}$(i,c) are correlated within clusters. As above, we focus on the asymptotic distribution of the GLS estimator, and we examine the case where one uses the true covariance matrix of the residuals within the level-two units. In this instance, the GLS estimator of β, β_{GLS}, is equivalent to the maximum likelihood estimator β_{MLE} . Using the same assumptions as above,

$$\sqrt{J}(\hat{b}_{MLE} - b) \xrightarrow{d} N(0, \Omega_{MLE})$$
where
$$\Omega_{MLE} = p\lim\left[\left(\frac{X'V^{-1}X}{J}\right)^{-1}\right].$$

V is the NxN covariance matrix of the residuals. Under the above assumptions it is given by

$$V = \begin{pmatrix} 1 & r & \cdots & r \\ r & 1 & \cdots & r \\ \vdots & \vdots & \vdots & \vdots \\ r & r & \cdots & 1 \end{pmatrix} \quad and$$

$$V^{-1} = \frac{1}{(1-r)[1+r(N-1)]} \begin{pmatrix} 1+r(N-2) & -r & \cdots & -r \\ -r & 1+r(N-2) & \cdots & -r \\ \vdots & \vdots & \vdots & \vdots \\ -r & -r & \cdots & 1+r(N-2) \end{pmatrix}$$

After some algebraic manipulation, the asymptotic covariance matrix, corresponding to the asymptotic variances and covariances of the three point estimators, can be expressed as

$$\Omega_{MLE} = \frac{1}{N} \begin{pmatrix} \frac{[1+r(N-1)][1+r((N-2)-t^2(N-1))]}{[1+r(N-2)](1-t^2)} & \frac{-t(1-r)[1+r(N-1)]}{[1+r(N-2)](1-t^2)} & 0 \\ \frac{-t(1-r)[1+r(N-1)]}{[1+r(N-2)](1-t^2)} & \frac{(1-r)[1+r(N-1)]}{[1+r(N-2)](1-t^2)} & 0 \\ 0 & 0 & \frac{(1-r)[1+r(N-1)]}{1+r(N-2)} \end{pmatrix} \quad \text{(A.4)}$$

The ratios of the variances for the OLS estimators, given by the diagonal elements of the covariance matrix in equation (A.3), to the corresponding variances of the maximum likelihood estimators defined in equation (A.4) provide a measure of the efficiency loss by using OLS instead of the efficient maximum likelihood estimator for the multilevel model. The square root of this variance ratio, which measures the ratio of the standard errors of the two estimators, provides an indication of the percentage increase in confidence intervals from using the less efficient OLS estimator. The efficiency loss due to using OLS instead of the maximum likelihood estimator for the multilevel model clearly depends on (1) the number of level-one observations within each of the J level-two units, (2) the intraclass correlation coefficient ρ, and (3) the degree of correlation between the community-level explanatory variable $X_C(c)$ and the individual level variable $X_{IC}(i,c)$.

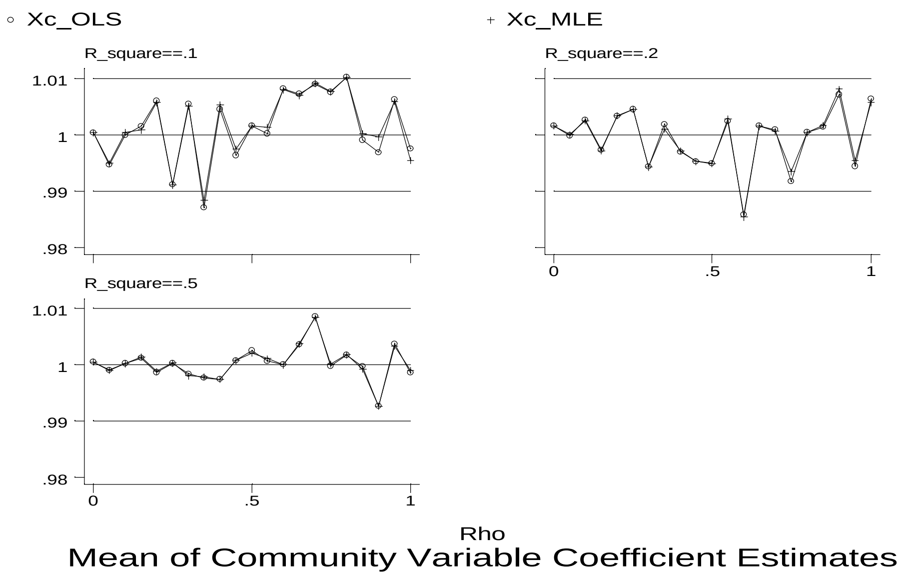
One exceptionally interesting variance comparison comes from the case where the community level covariates are uncorrelated with all of the individual level covariates (i.e., =0). In this instance, the variance ratio is 1 for two estimators of the impact the community level variable. This means that there is no efficiency loss in the estimation of the impact of the community level variable from using OLS instead of the maximum likelihood procedure when the community level variable is uncorrelated with community characteristics. One pertinent example when there would be a zero correlation is for the case where particular treatments (e.g., facilities or programs) are assigned randomly across communities. It is, however, important to note that the standard errors reported by a simple OLS procedure would be incorrect. The naive standard error reported by an OLS procedure that does not recognize the within level two unit error correlations would underestimate the true sampling variability of the OLS estimator by a factor of $[1+\rho(N-1)]$.[38]

[38]The incorrect, naive report of the asymptotic standard errors that a simple OLS procedure would yield are given by the square roots of the diagonal elements of the inverse of the E[(X'X)/J] matrix as presented in equation (A.2).

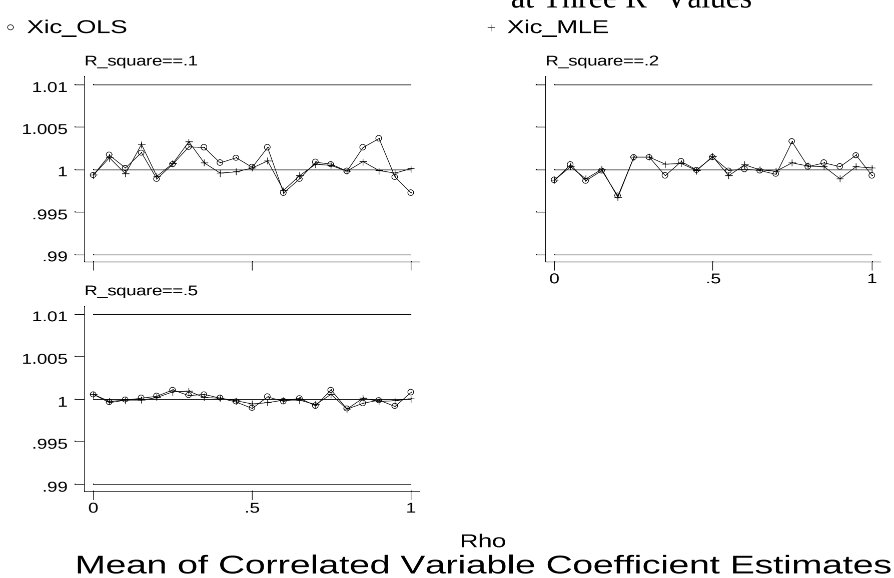
Appendix Figure 1

The Unbiasedness of Ordinary Least Squares and Maximum Likelihood Estimators in Models with Multilevel Errors by the Level of the Intraclass Error Correlation for 400 Communities, 50 Observations per Community

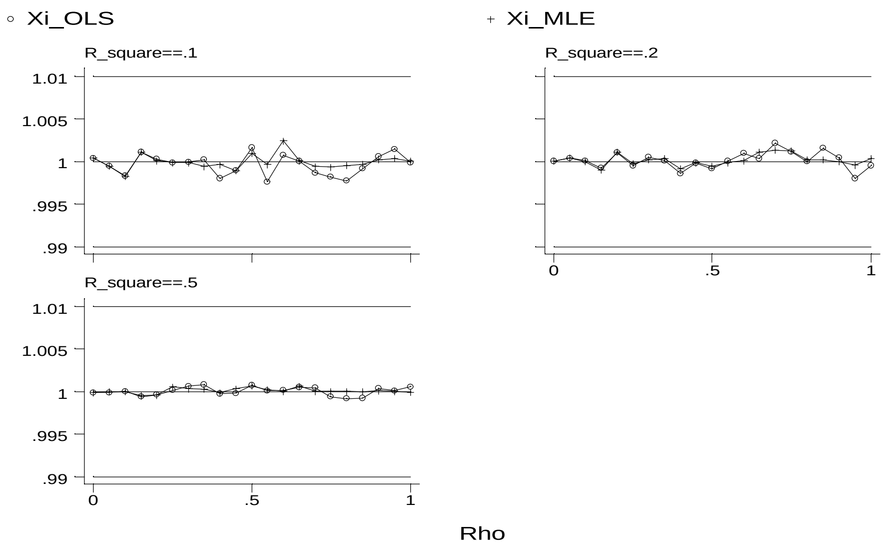
Appendix Figure 1 A: Community Level Variable Coefficient Estimates at Three R^2 Values



Appendix Figure 1 B: Correlated Individual Level Variable Coefficient Estimates at Three R^2 Values



Appendix Figure 1 C: Independent Individual Level Coefficient Estimates Three R^2 Values

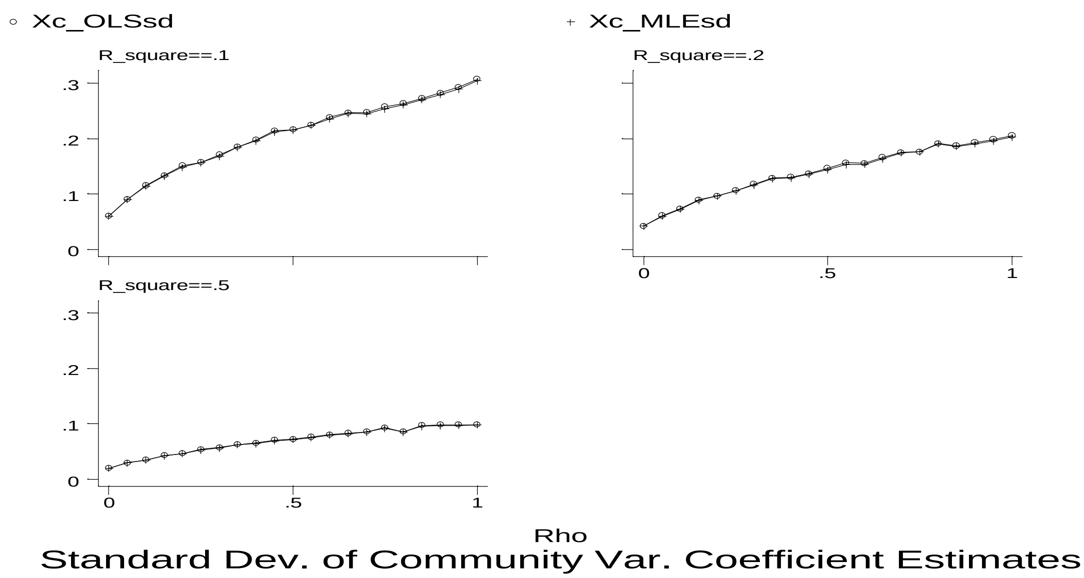


Mean of Independent Variable Coefficient Estimates

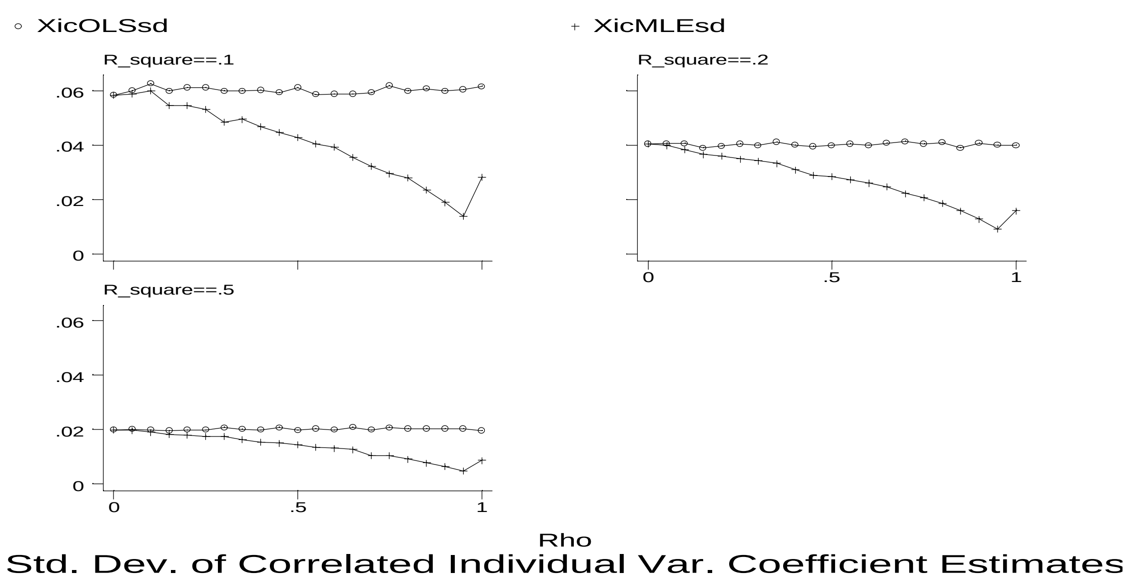
Appendix Figure 2

Empirical Standard Deviations of Ordinary Least Squares and Maximum Likelihood Estimates in Multilevel Models by the Level of the Intraclass Error Correlation for 400 Communities, 50 Observations per Community

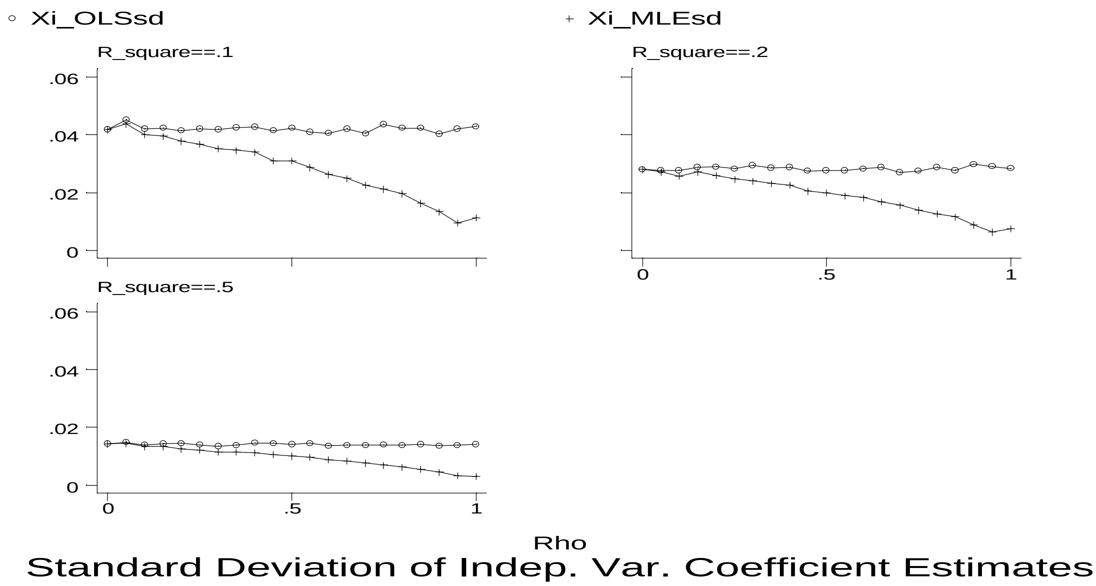
Appendix Figure 2A: Standard Deviations for Community Level Coefficient Estimates at Three R^2 Values



Appendix Figure 2B: Standard Deviations for Correlated Individual Level Coefficient Estimates at Three R^2 Values



Appendix Figure 2C: Standard Deviations for Independent Individual Level Coefficient Estimates at Three R^2 Values



References

Angeles G., Dietrich J., Guilkey D., Mancini D., Mroz T., Tsui A. and F. Zhang. 2001. "A Meta-Analysis of the Impact of Family Planning Programs on Fertility Preferences, Contraceptive Method Choice and Fertility." MEASURE Evaluation Working Paper Series No. 30.

Bollen, K., Guilkey D. and T. Mroz. 1995. "Binary Outcomes and Endogenous Explanatory Variables: Tests and Solutions with an Application to the Demand for Contraceptive Use in Tunisia," Demography: 32, 1.

Bryk A. and S. Raudenbush. 1992. Hierarchical Linear Models: Applications and Data Analysis Methods. Newbury Park: Sage

Eicker, F., 1963, "Asymptotic Normality and Consistency of Least Squares Estimators for Families of Linear Regressions," Annals of Mathematical Statistics, 24, 447-456.

Eicker, F., "Limit Theorems for Regressions with Unequal and Dependent Errors," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics, Vol. 1, Berkeley: University of California Press, pp. 59-82.

Gertler, P. and J. Molyneaux. 1994. "How Economic Development and Family Planning Programs Combined to reduce Indonesian Fertility," Demography: 31, 1.

Goldstein, H., 1995. Multilevel Statistical Models. London: E. Arnold; New York: Halsted Press.

Guilkey, D. and S. Cochrane. 1995. "The Effects of Fertility Intentions and Access to Services on Contraceptive Use in Tunisia," Economic Development and Cultural Change 43.

Guilkey, D. and S. Jayne. 1997. "Zimbabwe: Determinants of Contraceptive Use at the Leading Edge of Fertility Transitions in Sub-Saharan Africa," Population Studies 51.

Guo, G. and H. Zhao. 2000. "Multilevel Modeling for Binary Data," Annual Review of Sociology 26.

Huber, P. J., "The Behavior of Maximum Likelihood Estimates under Non-Standard Conditions," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics, Vol. 1, Berkeley: University of California Press, pp. 221-233.

Kalton, G., 1983, Introduction to Survey Sampling, Beverly Hills: Sage.

Kish, L., 1965, Survey Sampling, New York: John Wiley.

Kreft, I. and J. de Leeuw. 1998. Introducing Multilevel Modeling. Thousand Oaks, California: Sage.

Mroz, T., 2001, "Arbitrary Scaling and Spurious Biases in Multilevel Models for Discrete Outcomes," in progress, Carolina Population Center, UNC, Chapel Hill.

Rous, J., 2001. "Are Breastfeeding and Contraception Substitute Family Planning Strategies," forthcoming, Demography.

Stewart, J., Popkin B., Guilkey D., Akin J., Adair L. and W. Flieger. 1991. "Influences on the Extent of Breast-Feeding: A Prospective Study in the Philippines," Demography 28, 2.

White, H., 1980, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," Econometrica 48: pp. 817-830.